

This page ostensibly left blank.

Empirical Acquisition of Conceptual Distinctions via Dictionary Definitions

Thomas Paul O'Hara

May 18, 2004

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 for concepts relative to their sibling concepts. A two-step process is employed to decouple the definition parsing from the refinement 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 refine the relationship into their underlying semantics.

This empirical methodology is the main contribution of the thesis. In addition, a new probabilistic representation via Bayesian Networks is used for integrating this information with other statistical models. To illustrate the practical aspects of the work, the distinguishing information is used to improve a system for word-sense disambiguation.

Contents

| | | |
|---------|---|----|
| 1 | INTRODUCTION | 7 |
| 1.1 | Overview | 7 |
| 1.2 | Motivation | 8 |
| 1.2.1 | Differentiating relations are important | 8 |
| 1.2.1.1 | Support from Lexicography | 9 |
| 1.2.1.2 | Support from Psychology | 10 |
| 1.2.1.3 | Support from Knowledge Representation | 12 |
| 1.2.2 | Dictionary definitions are best source of differentiating relations | 13 |
| 1.3 | Contributions of this research | 14 |
| 1.3.1 | Empirical extraction and refinement of semantic relations | 15 |
| 1.3.2 | Representation of semantic relations using Bayesian networks | 15 |
| 1.3.3 | Improvements in word sense disambiguation | 16 |
| 1.4 | Organization of thesis | 16 |
| 2 | BACKGROUND ON LEXICAL SEMANTICS ACQUISITION | 17 |
| 2.1 | Background on Lexical Semantics | 17 |
| 2.1.1 | Linguistics | 17 |
| 2.1.2 | Lexicography | 20 |
| 2.1.3 | Computational Semantics | 20 |
| 2.1.3.1 | Semantic Networks | 20 |
| 2.1.3.2 | Word Experts/Agents | 21 |
| 2.1.3.3 | Ontological Semantics | 22 |

| | | |
|---------|---|----|
| 2.2 | Manual acquisition | 24 |
| 2.3 | Automated acquisition | 25 |
| 2.3.1 | Corpus analysis | 25 |
| 2.3.1.1 | Word Classes | 25 |
| 2.3.1.2 | Lexical Associations and Selectional Restrictions | 26 |
| 2.3.1.3 | Translation Lexicons | 28 |
| 2.3.2 | Lexical Rules | 29 |
| 2.3.3 | Analysis of dictionary definitions | 30 |
| 3 | DIFFERENTIA EXTRACTION | 36 |
| 3.1 | Analysis of definitions in WordNet | 37 |
| 3.1.1 | Structure of WordNet | 37 |
| 3.1.2 | WordNet definition annotations | 40 |
| 3.2 | Dependency parsing | 43 |
| 3.2.1 | Definition preprocessing | 43 |
| 3.2.2 | Parse postprocessing | 47 |
| 3.3 | Deriving lexical relations from the parses | 49 |
| 3.3.1 | Attachment resolution | 50 |
| 3.3.2 | Assigning relation weights using cue validities | 50 |
| 3.3.3 | Converting into nested lexical relation format | 51 |
| 3.4 | Differentia extraction algorithm | 52 |
| 3.5 | Issues | 52 |
| 4 | DIFFERENTIA REFINEMENT | 54 |

| | | |
|---------|---|----|
| 4.1 | Source and Target Term Disambiguation | 54 |
| 4.1.1 | Word-sense Disambiguation of Dictionary Definitions . . | 55 |
| 4.1.1.1 | Supervised WSD | 55 |
| 4.1.1.2 | Unsupervised WSD | 57 |
| 4.1.1.3 | Semi-supervised WSD | 57 |
| 4.1.2 | Using Sense Annotations from Extended WordNet . . . | 60 |
| 4.2 | Semantic Relation Inventories | 60 |
| 4.2.1 | Background on Semantic Roles | 62 |
| 4.2.2 | Inventories Developed for Corpus Annotation | 63 |
| 4.2.2.1 | Penn Treebank | 63 |
| 4.2.2.2 | FrameNet | 64 |
| 4.2.3 | Inventories from Knowledge Representation Frameworks | 65 |
| 4.2.3.1 | Cyc | 65 |
| 4.2.3.2 | Conceptual Graphs | 68 |
| 4.2.3.3 | Factotum | 70 |
| 4.3 | Relation Refinement | 73 |
| 4.3.1 | Overview of Relation Type Disambiguation | 73 |
| 4.3.1.1 | Use of Class-based Collocations | 75 |
| 4.3.1.2 | Classification Experiments | 76 |
| 4.3.2 | Penn Treebank | 77 |
| 4.3.2.1 | Illustration with ‘at’ | 79 |
| 4.3.2.2 | Results | 80 |
| 4.3.3 | FrameNet | 81 |

| | | |
|---------|---|-----|
| 4.3.3.1 | Illustration with ‘at’ | 82 |
| 4.3.3.2 | Results | 83 |
| 4.3.4 | Factotum | 84 |
| 4.3.4.1 | Inferring Semantic Role Markers | 86 |
| 4.3.4.2 | Method for Classifying the Functional Relations | 87 |
| 4.3.4.3 | Results | 87 |
| 4.3.5 | Combining the Different Semantic Role Inventories | 90 |
| 4.4 | Differentia Refinement Algorithm | 92 |
| 5 | APPLICATION AND EVALUATION | 95 |
| 5.1 | Lexicon Augmentation | 95 |
| 5.1.1 | Overview of Extracted Relations | 95 |
| 5.1.2 | Evaluation | 96 |
| 5.1.2.1 | Inter-coder Reliability Analysis | 98 |
| 5.1.2.2 | Results | 101 |
| 5.2 | Word Sense Disambiguation | 101 |
| 5.2.1 | Supervised Classification | 102 |
| 5.2.1.1 | Feature Overview | 102 |
| 5.2.1.2 | Differentia-based Features | 104 |
| 5.2.1.3 | System Results | 105 |
| 5.2.2 | Probabilistic Spreading Activation | 105 |
| 5.2.2.1 | Bayesian Network Representation | 105 |
| 5.2.2.2 | System Overview | 111 |
| 5.2.2.3 | System Results | 116 |

| | | |
|-------|--|-----|
| 6 | DISCUSSION AND FUTURE WORK | 120 |
| 6.1 | Related Work | 120 |
| 6.1.1 | Differentia Extraction | 120 |
| 6.1.2 | Relation Refinement | 122 |
| 6.1.3 | Semantic Relatedness | 124 |
| 6.1.4 | Relation Weighting | 124 |
| 6.1.5 | Word Sense Disambiguation | 125 |
| 6.1.6 | Class-based Collocations | 126 |
| 6.1.7 | Bayesian Networks | 127 |
| 6.2 | Future Work | 128 |
| 6.2.1 | Application to Text Segmentation | 128 |
| 6.2.2 | Mapping Senses from other Dictionaries into WordNet . | 130 |
| 6.2.3 | Transferring Semantic Roles across Resources | 130 |
| 6.2.4 | Inferring other Types of Relations | 131 |
| 6.2.5 | Analyzing Lexical Gaps | 133 |
| 7 | CONCLUSION | 134 |
| 7.1 | Summary of Thesis | 134 |
| 7.1.1 | Importance of Differentiating Relationships | 134 |
| 7.1.2 | Approaches for Lexical Acquisition | 135 |
| 7.1.3 | Extraction of Differentiating Relations | 135 |
| 7.1.4 | Refinement into Conceptual Relations | 136 |
| 7.1.5 | Lexicon Augmentation and Word-sense Disambiguation | 136 |
| 7.1.6 | Looking Backward and then Forward | 137 |

| | | |
|-------|---|-----|
| 7.2 | Significance of Research | 138 |
| 7.2.1 | Empirical Acquisition of Distinctions from Dictionaries . . | 138 |
| 7.2.2 | Exploiting Resources on Relation Usage | 138 |
| 7.2.3 | Bayesian Networks for Differentia Representation | 138 |
| 7.2.4 | Class-based Collocations for Word-Sense Disambiguation | 139 |
| 7.3 | Speculations regarding Computational Semantics | 139 |
| | APPENDICES | 141 |
| A | PRIMER ON BAYESIAN NETWORKS | 142 |
| B | PRIMER ON MACHINE LEARNING | 145 |
| | REFERENCES | 152 |

CHAPTER 1 INTRODUCTION

This thesis aims to improve semantic lexicons for computational linguistics by automatically extracting information from conventional dictionary definitions (e.g., English language definitions). The motivation for the work is that broad-coverage lexicons often do not provide sufficient information to differentiate sibling concepts. Consequently, different words mapping to such undifferentiated siblings concepts are effectively treated as synonymous. 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.

Although this problem has been addressed by various approaches in the past, most of the previous work has relied upon manually derived extraction rules, which makes it more difficult to adapt 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 refinement 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.

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 computational linguistics, as well as from other disciplines, namely psychology and lexicography. Section 1.3 then presents more details on the contributions of the thesis. Lastly, Section 1.4 outlines the rest of the thesis.

1.1 Overview

Computational lexicons are good at conveying the main semantic 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, 'beagle'

and ‘wolfhound’ are considered equivalent with respect to the explicit WordNet relations that apply to both. There is no indication of the size attributes that clearly distinguish the two concepts. This problem also exists to a lesser extent with traditional knowledge bases, of which computational lexicons are a special case.

Dictionaries are a good source of these relations. In fact, dictionaries have evolved into repositories of important distinguishing characteristics for words similar in meaning. Several reasons account for this aspect of definitions, such as the perception that dictionaries are authorities on language use as well as publishing constraints.

This thesis presents an approach for extracting differentiating information from dictionary definitions, also known as *differentia*. Unlike previous approaches, it 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 programming. Instead, an empirical approach addresses this problem by requiring additional annotations. This in effect replaces the programming requirement of knowledge-based approaches with an annotation requirement. 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. For instance, it is possible to infer the annotations from existing knowledge resources.

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 provide a more direct means of obtaining this infor-

mation 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). But nonetheless 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*:¹

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 is too different 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 meaning of an unknown word or of an unfamiliar sense of a known word. The first aspect (word choice) is addressed by having the dictionary concentrate more

¹From <http://stage1.worldbook.com/products/htmla/mw.htm>.

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 the 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 *differetiae*), again making the analytic definition suitable. For instance, consider the LDOCE² definition of ‘money’ versus the corresponding one in WordNet³

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. But this requires the use of additional differentiating information, for instance, to distinguish money from other “pieces of metal.”

Also note that the dictionary usually just *relates* a word (or word sense) to an underlying (superordinate) concept and provides 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. So again the emphasis 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:

²Longman’s Dictionary of Contemporary English (Procter, 1978)

³Version 1.7 of WordNet is used throughout the discussion.

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

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 is able to account 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, it is the distinctive features of the italicized terms that set the stage for the analogy. In addition, Tversky later conducted experiments (Gati and Tversky, 1984) that showed 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 (1975) research into the use of family resemblances in categorization highlighted the use of distinctive features as well as common features in categorization. One important finding is that natural categories⁴ are generally organized so as to maximize the similarity within a class and 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. In fact, the notion of family resemblances can be cast in terms of cue validities, although she prefers the former term for clarity.

⁴Natural categories are those for which people can readily associate typical members. For instance, *chair* would be a natural category instead of *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 information resides, in terms of attributes (Rosch, 1973; Rosch and Mervis, 1975).

1.2.1.3 Support from Knowledge Representation

Conceptual knowledge is commonly organized into hierarchies that are called ontologies (Mahesh and Nirenburg, 1995). The concepts in these hierarchies are usually partially ordered via the subset relation (*is-a*). This is the *relation of dominance* that Cruse (1986) gives 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 that a word is defined relative to. 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, Mikrokosmos⁵ averages about 2.4 properties per concept (including some inverse relations), whereas WordNet⁶ only averages 1.3 (including inverses).⁷

This suggests that the reason large-scale lexicons tend not to include such differentiating relations is due to 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.

⁵1998 version of Mikrokosmos (crl.nmsu.edu/Research/Projects/mikro/index.html).

⁶Version 1.7 of WordNet (www.cogsci.princeton.edu/~wn).

⁷*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 best source of differentiating relations

As mentioned earlier, corpus analysis is unlikely to be a good source of differentiating relations. For example, collocations⁸ for a particular word (or word sense) might include words that indicate aspects covered by the differentiating relations. But 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. The following words showed up when using a standard conditional independence test (Wiebe et al., 1997) for selecting indicative collocations for each sense selected from the Brown portion of the DSO⁹.

city#2: an incorporated administrative district established by state charter

\$ 1 15,000 30 31 Arthur California Carolina Coosa Dakota December Georgia Hemphill Hughes July June Levitt North Portland Providence Rhode September Sokol Virginia *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#1: the territory occupied by one of the constituent administrative districts of a nation

& - 11 12 13 1960 1963 23,000,000 25 8 95 Connecticut Illinois Indiana Massachusetts O Parker Vermont Warwick Washington adjust adjustment air allot allotment along assess assessment

⁸Collocations are used here in the broader sense of words that co-occur often in context: there are no constraints on word order, etc.

⁹Singapore’s Defense Sciences Organization corpus (Ng and Lee, 1996)

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 four-
teen 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 re-
spective respectively secede sell snow southeastern stay sum
tangible taxation team thereby thirteen twenty-five unadjusted uni-
formly upward

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 have *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#2* (or vice versus), although a few states are listed along with other proper names. The closest connection would be through 'municipal,' listed under the collocations for *state#1*. 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, these 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 it is given in natural language, which leads to the usual problems with structural and lexical ambiguities during analysis. But 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.

1.3 Contributions of this research

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

1.3.1 Empirical extraction and refinement of semantic relations

Most earlier approaches relied upon manually constructed 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 refined into semantic relations between the underlying concepts. In other words, the surface-level syntactic relationships determined by the parser are refined into the underlying conceptual relations. Standard word-sense disambiguation (WSD) approaches are used to refine the terms being related; this aspect is not a focus of the research, given the recent advances in WSD. Statistical classifiers are used to refine the relation types, exploiting both tagged corpora and knowledge bases for the training data. Isolating the refinement 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 refinement process.

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). 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 end 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. The original approach combined analytical knowledge about the dependencies among word senses taken from 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 rest of the thesis proper is organized as follows: Chapter 2 presents related work mainly in lexical acquisition but also covers work in computational linguistics 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). Chapter 4 discusses the refinement of the surface-level relationships into conceptual relationships. This discusses the disambiguation of the source and target terms and the refinement of relation types from their English specification (e.g., preposition) into the underlying concept for the relation type.

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 computational linguistics (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 information 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. Lastly, 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 gives a basic introduction to Bayesian networks, which are popular for representing probabilistic relations in artificial intelligence. Appendix B explains the general framework for machine learning and discusses the two main types used in this research.

CHAPTER 2 BACKGROUND ON LEXICAL SEMANTICS ACQUISITION

This thesis approaches the task of conceptual refinement 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, as well as providing a background in the representation and utilization of such lexical information.

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 tradition (Dowty, 1979) accounts more centrally for the semantics of words, such

| 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 |
| factive | 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).

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 and that surface case relations are often insignificant in that 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 would include, although more would be needed in practice (e.g., benefactive). See Table 2.1.

Jackendoff (1983) presents a unified framework for representing conceptual knowledge for all aspects of cognition, not just linguistic. 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 has a few innovations that are noteworthy. 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 his 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 TO_{+dist} [place ON_{+dist} [wall]]])].

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

| Relation | Description |
|--------------------|---|
| hyponymy | $z \text{ in } X \Rightarrow z \text{ in } Y$ |
| taxonomy | X is a kind/type of Y |
| meronymy | X is part of Y; also called <i>partonymy</i> |
| 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).

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. 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 about the Levin verb classes (1993). 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 context, whereas syntagmatic relations hold between elements that can occur together in the same 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 se-

mantic 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 good insight into lexical semantics, in particular regarding word-sense distinctions. Kilgarriff (1997) calls to task the notion of there being a distinct set of word senses independent of their use in particular application, which is commonly assumed by word-sense disambiguation (WSD) researchers. Support for this claim is that what appear as separate senses in definitions are the result of lexicographers distilling disparate usages from citation files. As not all the citations can be addressed, inevitably some senses will not be accounted for in dictionaries.

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 like '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. Furthermore, dictionaries often do not clearly indicate that such encyclopedic information is used 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. Also, it was one of the first computational attempts to emphasize semantics over syntax. His work centered around 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. Furthermore, 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 vs. 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 (1982) popularized the idea of *word experts* 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 (1994) illustrate 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 done by establishing variable linkages between argument placeholders in the syntactic structure (SYN-STRUCT) and concept placeholders in the semantic structure (SEM-STRUCT). 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 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, such as to allow 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 a variety of lexical assertions. Proper name assertions map strings to individuals in the KB. A denotational assertion

```

(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")
    (syn)                  ;; syntactic features
    (syn-struct            ;; syntactic structure
      (1 ((root $var0)
          (cat n)) ))
    (sem-struct            ;; 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'.*

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 (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 (e.g., *is-a* and *has-a*). Thus, WordNet represents an implicit ontology in which the concepts are lexicalized. 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 ma-

nipulated 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 performativeness.

2.2 Manual acquisition

Manual acquisition has been most commonly used when the lexicon quality is critical (Onyshkevych and Nirenburg, 1994). 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 partially 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 which 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. 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, both due to increased ambiguity as WordNet got larger and to requests from users who expected fuller definitions. Recently, there has been some 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, and even when so undesirable, unless the quality of the information is guaranteed to be at the same level of 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 as their similarity measure:¹

$$p_n(v) = f_{vn} / \sum_v f_{vn}$$

$$D(p_{n1} || p_{n2}) = \sum_c p_{n1}(x) \log\left(\frac{p_{n1}(x)}{p_{n2}(x)}\right)$$

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

¹Entropy is defined as $\sum_x -p(x)\log(p(x))$. See Appendix B for more details.

is first used to extract dependency tuples of the form $\langle \text{word1}, \text{grammatical-relation}, \text{word2} \rangle$. He measures word similarity based on frequency of the tuples and their constituents using mutual information (Manning and Schütze, 1999), which measures difference in the joint occurrence of two events versus the occurrence expected by chance (i.e., $-\log(p(xy)/p(x)p(y))$). The mutual information (MI) for the co-occurrences of two words in a particular grammatical relationship defined as follows:

$$MI(w1, r, w2) = -\log((P(r)P(w1|r)P(w2|r))/P(w1, r, w2))$$

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

Slator et al. (Slator et al., 1990) apply clustering to preposition descriptions derived from LDOCE in order to automatically derive 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 somewhat 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 triples of the form $\langle \text{verb}, \text{noun}, \text{prep} \rangle$ 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 $\langle \text{verb}, \text{prep} \rangle$ and $\langle \text{noun}, \text{prep} \rangle$ 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 triples. Then, instead of using the heuristics on each ambiguous triple (i.e., those with both verb and noun non-empty), the bigram frequencies are used in a log-likelihood ratio test:

$$\log_2 \frac{P(\text{verb_attach } p \mid v, n)}{P(\text{noun_attach } p \mid v, n)}$$

where $P(\text{verb_attach } p \mid v, n)$ is estimated by $\text{freq}(\text{verb, prep})/\text{TotalFreq}$ 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. They define *semantic expectation* as the probability that a pair of concepts occurs in a given relationship. 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, that is, the determination of the selection restrictions for particular words. An experiment in deciding prepositional attachment shows how this method improves over 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 cue validities discussed in the last chapter (see Section 1.2.1.2):

$$\sum_{k=1}^K P(C_k) \sum_{ij} P(\text{attr}_i = \text{val}_j \mid C_k)^2$$

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 triples. 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 is done to avoid the data sparsity problem associated with statistical inference at the word level. A benefit of doing this is that the classes 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*, which 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 (Resnik, 1993):

$$S(p) = D(\Pr(C|v)||\Pr(C)) = \sum_c \Pr(c|p) \log(\Pr(c|p)/\Pr(c))$$

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

$$A(p, C) = \frac{\Pr(C|p) \log\left(\frac{\Pr(C|p)}{\Pr(C)}\right)}{S(p)}$$

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

There has been a good deal of 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., 1992) implemented steps to partly automate this process, such as in the use of semantic dictionaries inferred from the training data. Riloff and Schmelzenbach (1998) illustrate further automation of this process in 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 examples 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 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 fix 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 pig 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. An example of this would be the derivation of ‘comprador’ (buyer) from ‘comprar’ (to buy). They also point out difficulties associated with lexical rules, such as for English adjectives. For example, although ‘-able’ is a very productive affix for converting a verb into an adjective, 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 of 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 basic one. 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

In the mid-80’s, a trend began towards building more realistic applications of natural language processing. Earlier work, in addition to being restricted to toy 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. This involved the extraction of the genus headwords along the lines of Amsler’s (1980) manual analysis of Webster’s pocket dictionary. Much work was done with Longman’s Dictionary of Contemporary English (LDOCE), both because of favorable research licensing and due to its restricted vocabulary of defining terms (Procter, 1978). (Boguraev and Briscoe, 1989) contains a good survey of early LDOCE-related research. It also illustrates some of the difficulties commonly encountered, such as in dealing with errors in the typesetting formatting and inconsistencies in the definitions.

Later on, additional implicit information was extracted from the dictionaries. Alshawi (1989) and Jensen and Binot (1987) applied pattern matchers to extract case or thematic relations among words defined in the entries. There have been similar approaches applied since, most recently by Barrière (1997) and Vanderwende (1994; 1996). See (Wilks et al., 1996) for a comprehensive survey of work on MRD-related research.

The main contribution of Amsler’s (1980) thesis is the development of procedures for the extraction of genus hierarchies from machine readable dictionaries (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, such as in the analysis of the differentia descriptors used in motion verbs, suggestions for parsing dictionary definitions, indications of what might be expected from a deep analysis (unraveling mor-

phological relations), and the description of 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 6.1.5. 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, such as the following, 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) describe 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. This extraction is in support of a system for resolving prepositional attachment. To decide on an attachment they first check whether the complement's definition has a pattern indicative of one of the relations considered and if so whether linkages can be established to the head given the hierarchy of the definition genus terms.

Wilks et al. (1989) describe three different methods for analyzing LDOCE. All deal with aspects of taking the existing information in LDOCE and converting into *machine tractable dictionaries* (e.g., lexical knowledge bases). The first method is based on co-occurrence analysis of the control 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 (grammar code, 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.

In work related to the above, Slator and Wilks (1987) sketch out an approach for deriving rich lexical entries from the information present in LDOCE. This first incorporates the explicit information available in the online version of LDOCE: the extended grammatical category, semantic restrictions, and pragmatics code. The definitions are then parsed to extract information from the differentia. 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 described 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 defining 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 yield 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 details semantic information from LDOCE, building upon the work of Jensen and Binot (1987) that uses a general-purpose parser rather than string matching. Subcategorizations are only used to guide the parse, not rule out potential parses, which is important because definitions incorporate various forms of ellipsis more than normal text; and, a single parse is always produced (using the default right attachment but indicating other potential attachment sites). 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

| | |
|----------------|---------------|
| 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)

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 thesis (Barrière, 1997) 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; refining the lexical relations in the conceptual graphs; and, combining the conceptual graphs from different definitions. The grammar is a simple context free grammar with some customization to dictionary definitions, such as the use of the “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 of the shallow parses for the MRD 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

| | | |
|----------------|---------------|--------------|
| 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)

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 upon of LDOCE’s control 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 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
⇒ [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:

NP → NP VP ⇒ noun(obj2) & verb(e, subj1, obj2).

CHAPTER 3 DIFFERENTIA EXTRACTION

This thesis is motivated by the desire to have finer conceptual distinctions in semantic lexicons and in knowledge bases in general. To this end, the distinguishing relations (*differentia*) indicated in dictionary definitions are extracted and used to augment WordNet, a common lexical resource for natural language processing.

An empirical approach to acquiring finer conceptual distinctions is taken here. To acquire the information from the definitions, a general-purpose parser combined with example-based learning is used rather than manually encoding pattern matching rules. Furthermore, the acquisition process is split into two parts: extraction of lexical relations and refinement into conception relations. The first step (*Extraction*) involves determining the important surface-level relations present in the definitions. A broad coverage dependency parser¹ is used for this aspect. The second step (*Refinement*) involves determining the underlying conceptual relations indicated by the lexical relations. This is based on automatic semantic roles classifiers inferred using machine learning over manual annotations, as discussed in the next chapter.

This chapter is organized as follows. Section 3.1 presents an overview of WordNet and discusses manual annotations of a subset of the definitions. Section 3.2 discusses the parsing proper. Definitions taken from a dictionary are first preprocessed to make more suitable for parsing (e.g., conversion to complete sentences.) Then the sentences are parsed producing a list of low-level syntactic relations among the words. Section 3.3 discusses the initial relation extraction. Postprocessing is done to convert these into more traditional grammatical relations and to make the relations centered around function words in preparation for the refinement step. In addition, grammatical relations are weighted based on the degree of specificity to particular concepts and then the relations are converted into a more readable format. Section 3.4 summarizes the extraction algorithm; and, Section 3.5 closes with a discussion of issues involved in the extraction process.

¹Dependency parsers stress the links between words rather than the structural configuration of syntactic categories, which is typical of traditional phrase structure parsers (Sleator and Temperley, 1993).

3.1 Analysis of definitions in WordNet

The dictionary used here as the source of definitions is WordNet (Fellbaum, 1998).² WordNet incorporates aspects of a thesaurus as well as a dictionary. Words are grouped into synonyms 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}.

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 '⇒' links in the figure are for WordNet's *hypernym* relation (same as *is-a*). Table 3.2 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 Longman's Dictionary of Contemporary English (Procter, 1978), although not as comprehensive as a college dictionary such as Merriam Webster's Collegiate Dictionary (Mish, 1996). It covers the core lexicon of English, but also includes some scientific and technical terms. Table 3.1 shows some statistics on the number of entries in WordNet. The *Entries* column is 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* refers to the number of total number of sense distinctions for all the entries (e.g., six for 'dog'). These are numbered as often done in traditionally dictionaries, but there are no further subdivisions (e.g., '1a'). In contrast, *Synsets* refers to the number of underlying concepts, the targets for the senses. Unlike traditional dictionaries, definitions are not given for the word senses but instead for the synsets. This is not apparent from Figure 3.1, but can be inferred from Figure 3.2 which shares the following synset:

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

That is sense 4 of 'dog' and sense 1 of 'cad' refer to the same underlying synset.

²WordNet version 1.7.1 is used throughout unless noted elsewhere. WordNet is freely available from Princeton. The database along with full documentation can be found at www.cogsci.princeton.edu/~wn.

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: **WordNet definitions for the noun ‘cad’ in WordNet.**

| 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 3.1: **Number of entries in WordNet 1.7.1.**

| 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 3.2: ***Relation usage in WordNet (version 1.7).***

3.1.2 WordNet definition annotations

Manual annotations are commonly used in computational linguistics to provide insight of the genre of text being studied. They also are used to develop 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 as occurring in parts of the Wall Street Journal and Brown Corpus, yielding 190,000 distinct annotations.

To provide insight into the WordNet definitions, annotations of the semantic relations implicit in the definitions were done for the subset of the definitions corresponding to words sense-annotated in the DSO corpus. (Future work will look into using the annotations as input into the differentia extraction process as discussed in Chapter 6). 170 of the words were randomly selected as training data, with the remaining 20 set aside to be used for held-out test data. Roughly 650 of 1500 definitions from the training data were annotated, yielding over 2400 tagged relation occurrences. Figure 3.3 shows a sample of the annotations.³ Table 3.3 shows the most common semantic relations used in

³See www.cs.nmsu.edu/~tomohara/wordnet-definition-annotations for the complete set of annotations.

| Relation | Frequency | Description |
|--------------|-----------|---|
| genus | 632 | category for concept (definition <i>genus</i>) |
| attr | 458 | generic attribute |
| object | 306 | affected object or event |
| spec | 247 | specialization involving given type of participant |
| manner | 109 | manner in which action is done |
| qual | 84 | qualification of the genus category |
| subject | 60 | generic actor for action described |
| example | 56 | example for which the definition applies |
| alt-genus | 55 | secondary category for concept |
| location | 40 | physical or abstract location |
| purpose | 37 | what an object is used for or why an action is done |
| result | 23 | result produced by some action |
| means | 20 | means by which an action or condition is achieved |
| concerning | 14 | objective theme used for specialization |
| subject-attr | 14 | attribute of sentence subject (actor) |
| action | 13 | action in descriptive subordinating clause |
| agent | 12 | agent performing an event |
| field | 12 | domain indicated by usage label |
| source | 11 | physical or abstract starting point |
| complement | 11 | non-objective complement of a verb |
| time | 9 | point in time or relative time of occurrence |
| part-of | 8 | constituent of another entity or action |
| restriction | 8 | specialization to a particular domain area |
| destination | 8 | physical or abstract ending point |
| pertains-to | 7 | specialization involving given topic area |

Table 3.3: ***Common semantic relations from the definition annotations.***

these annotations, along with a description of each. The next chapter will discuss other semantic role annotations, mainly those centered around thematic roles. Thematic roles are included here (e.g., *location* and *source*); however, specialization-type relations are more commonly used (e.g., *spec*, *qual*, and *concerning*). This reflects the focus of definitions on definite descriptions rather than actions or situations in general. Furthermore, the emphasis of specializations adds support that definitions are differential in nature.

Nouns:

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

law#1: the collection of rules imposed by authority

- genus collection
- spec rules
- qual imposed
- agent authority

Verbs:

add#4: make an addition by combining numbers

- genus make
- object addition
- manner combining
- object numbers

draw#31: remove the entrails of

- genus remove
- object entrails
- spec <vp_object>

keep#15: maintain by writing regular records

- genus maintain
- means writing
- object records
- attr regular

Figure 3.3: ***Sample of WordNet semantic relation annotations.***

3.2 Dependency parsing

The approach to differentia extraction is entirely automated. This starts with using the *Link Grammar Parser* (Sleator and Temperley, 1993), a dependency 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.

3.2.1 Definition preprocessing

Dictionary definitions are often given in the form of sentence fragments with the headword omitted. Some learner's dictionaries now give definitions in complete sentences (Barnbrook, 2002), but this is not yet common practice. For example, the definition for *lock*_{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. Fortunately, definitions for words of a given part of speech are usually given in 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 headword in a given sentence without affecting grammaticality (Landau, 2001).

At least 80% of the noun definitions in WordNet follow this pattern, based on tabulations from running a part of speech tagger and then a simple phrase chunker. Table 3.4 shows the top part of speech sequences occurring in the noun definitions. A similar analysis for the verb definitions show 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 assignments by the part-of-speech tagger. These tend to have trouble tagging fragments because the complete sentential context is not available. Also note that the WordNet definitions have not been through the amount of editing that definitions for commercial dictionaries go through.⁴ Therefore, the latter undoubtedly will have higher percentages as a result of better quality control (e.g., more uniformity).

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 extraneous semantic content introduced. For nouns, the optional

⁴The WordNet definitions were initially not included in the lexicon, because it was felt the synset groupings would help 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 verb NP | 624 |
| NP verb to-preposition verb NP | 331 |
| NP punctuation verb NP | 266 |
| NP verb preposition NP punctuation | 261 |
| NP adjective preposition NP | 257 |
| NP punctuation verb preposition NP | 254 |

Table 3.4: ***Top patterns for WordNet noun definitions (76,191 total).***

| 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 |
| verb NP preposition | 111 |
| verb NP to-preposition punctuation | 100 |
| verb adverb | 97 |
| verb adjective conjunction adjective punctuation | 92 |
| verb preposition NP punctuation preposition preposition NP | 87 |

Table 3.5: ***Top patterns for WordNet verb definitions (13,406 total).***

| 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.**

determiner is used whenever the word is a count noun (e.g., ‘dog’ in contrast to ‘sand’). Also, the determiner ‘a’ is used unless the word being defined starts with a vowel. Thus, for *lock*_{fastener}, the following sentence would be used for the parse: “A lock is a *fastener fitted to a door or drawer to keep it firmly closed.*” In contrast, for *anemia*_{disease}, the result would be “Anemia is a *deficiency of red blood cells.*”

For verbs, the definition is first changed if necessary to be an infinite phrase. The definition is then used as a verbal complement to a subject infinitive phrase formed from the headword. For *delay*_{pause} this yields the definitional sentence “To delay is to *act later than planned, scheduled, or required.*” A better pattern might be to just convert the definition into the past tense and supply a dummy subject of ‘it’ (e.g., “It *acted later than planned, scheduled, or required.*”). This would make it easier to later 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, perhaps incorporating different patterns based on the type of 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 sequence patterns for these 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 preposition phrases, with a special case being ‘in a <adjective> manner’ (779 cases). Although most cases use simple defining phrases, occasionally these are defined in terms of a series of items indicating various aspects suggested by the term. 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 disjunctions. 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 “The incredible thing is *beyond belief or un-*

| 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 |
| adverb verb punctuation | 162 |
| verb conjunction verb preposition NP punctuation | 151 |
| preposition NP | 130 |
| verb conjunction verb NP punctuation | 111 |
| verb to-preposition NP punctuation | 111 |

Table 3.7: ***Top patterns for WordNet adjective definitions (18,700 total).***

| 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 |
| preposition determiner verb NP | 20 |
| preposition conjunction preposition NP punctuation | 20 |
| preposition verb NP punctuation | 20 |
| adverb conjunction adverb punctuation | 19 |
| preposition NP conjunction preposition NP punctuation | 17 |

Table 3.8: ***Top patterns for WordNet adverb definitions (3,636 total).***

derstanding.” 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.*”

An alternative approach to converting the definitions into sentences would be to customize the grammar of the parser to accommodate the definition fragments. However, this would involve quite a bit of work, much of which might be specific to the parser being used.

As preprocessing would nonetheless still be required, it is better to do more preprocessing since that is less dependent on the parser implementation.

3.2.2 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 outputs syntactic dependencies among words, punctuation or dummy elements. For example, Figure 3.4 shows the raw parse for the definition of ‘wine’ given as “fermented juice (of grapes especially).” The parser uses quite specialized syntactic relations, so these are converted into general ones prior to the extraction of the relational tuples. For example the relation *A*, which is used for pre-noun adjectives, is converted into *modifies*. Table 3.9 shows the conversions for the some of the common relations encountered.⁵ 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 dummy elements or punctuation. For example, the first tuple extracted from the parse would be ⟨n:wine, *v:is*, n:juice⟩. Certain types of dependencies are treated specially by converting the syntactic

⁵Detailed documentation on the Link Grammar Parser relation types can be found at <http://hyper.link.cs.cmu.edu/link/dict/summarize-links.html>.

| Relation | Replacement | 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 |
| Cr | subject-of | subject of relative clause |
| Cs | subject-of | subject of subordinated clauses |
| 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 |
| SF_ | subject-of | filler subjects (e.g., 'it') to finite verbs |
| TO_ | modal | verbs and adjectives to the word 'to' |
| W_ | to-wall | marks beginning or end of sentence (the <i>wall</i>) |
| X_ | is-punctuation | used with punctuation |
| YS | is-possessive | nouns to the possessive suffix |

Table 3.9: **Sample mapping from Link Parser relation types into general ones.** 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).

definition sentence:

Wine is fermented juice (of grapes especially).

parse:

```
<///// , Wd, 1. n:wine>
<///// , Xp, 10. .>
<1. n:wine, Ss, 2. v:is>
<10. . , RW, 11. /////>
<2. v:is, Ost, 4. n:juice>
<3. v:fermented, A, 4. n:juice>
<4. n:juice, MXs, 6. of>
<5. (, Xd, 6. of>
<6. of, Jp, 7. n:grapes>
<6. of, Xc, 9. )>
```

Figure 3.4: **Link Grammar parse for wine**_{alcohol}.

```
<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**_{alcohol}.

relationships directly into a relational tuple involving a special relation-indicating word (e.g., ‘modifies’). For the example, this yields the tuple ⟨v:fermented, *modifies*, n:juice⟩. 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 refinement).

The result of the initial conversion for the wine example is shown in Figure 3.5.

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. The remaining steps in the process account for structural ambiguity in the parses and also for assigning weights to the relations that are extracted.

Cue validity of feature F for concept C :

$$\begin{aligned} P(C|F) &= \frac{P(F|C)}{\sum P(F|C_i)} \\ &= \frac{f(F,C)/f(C)}{\sum f(F,C_i)/f(C_i)} \end{aligned}$$

where C_i is a concept that contrasts with C

Figure 3.6: **Calculation of cue validities.**

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 handling 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 which 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 relations in applications, it is desirable to have a measure of how relevant the relations are to the associated concepts. One such measure would be the degree to which the relation applies to the concept being described as opposed to sibling concepts. To account for this, *cue validities* are used. As discussed in the previous chapter, these can be interpreted as probabilities indicating the degree to which features apply to a given concept versus similar concepts.

Cue validities are estimated by calculating the percentage of times that the feature is associated with a concept versus the total associations of contrasting concepts, as shown in Figure 3.6. This requires a means of determining the set of contrasting concepts for a given concept. The simplest way of doing this would be to just 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 con-

cepts 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*. Furthermore, the related concept *lecturer*_{educator} would not be considered as contrasting with either since its parent is *educator*, which is the parent of *teacher* and thus is neither a sibling nor cousin concept to either of the above.⁶

To alleviate this problem the *most-informative ancestor* will be used instead of the parent. This is determined by selecting the ancestor that best balances frequency of occurrence in a tagged corpus 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 most information content). Here attested frequencies from SemCor 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 technical.

The cue validities for all of the lexical 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.) First, for each concept associated with lexical relations (e.g., the concept for the definition headword), its most-informative ancestor (MIA) is determined. Associations are then updated for each of the features present in the lexical relations, (i.e., $f(F, C)$ in Figure 3.6). 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., $\sum f(F, C_i)$).

3.3.3 Converting into nested lexical relation format

The unstructured relation listing resulting from the parse postprocessing can be unreadable when the parses are complex. To alleviate this problem, the relations are converted into a format that incorporates nesting to account for the interrelationships. In addition, the offsets are removed since the nesting accounts for it. This makes it easier for humans to evaluate the results and facilitates revisions in case humans will be doing post-editing of the relation

⁶Subscripted category names (e.g., *lecturer*_{educator}) are used instead of sense indicators (*lecturer#1*) when referring to synsets by words that are ambiguous. An alternative would be to list all the words in the synset (e.g., {lector, lecturer, reader}), but that becomes awkward to use in text.

```

sense: noun:artifact#1

definition sentence: An artifact is a man-made object.

relational tuple format:

<1. a, 2. n:artifact, 3. v:is>
<2. n:artifact.n, 3. v:is, 6. n:object>
<5. man-made, modifies, 6. n:object>

nested relation format:

noun:artifact\#1:
    verb:is n:object
        attribute man-made

```

Figure 3.7: ***Example conversion of flat relation tuples into nested relation format.***

extraction prior to incorporation in a lexicon. Figure 3.7 shows an example of this conversion.

3.4 Differentia extraction algorithm

Figure 3.8 presents a high-level description of the differentia extraction algorithm. The code has been implemented using Perl. It is available for download at www.cs.nmsu.edu/~tomohara/differentia-extraction.

3.5 Issues

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 properties that are being extracted are indeed differential. This is not a focus of the 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. However, the

1. Preprocess definitions, 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 syntactic to canonical format with higher-level syntactic relations with relation types based on function words
5. Disambiguate the parses (see Chapter 4).
6. Weight the relations based on cue validities
7. Convert from flat relational tuples into nested lexical-relation format

Figure 3.8: ***Differentia extraction algorithm.***

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 property is would be to compare its cue-validity weight versus those for other properties. Because this will have a bias towards incidental co-occurrences (e.g., greyhound \iff bus line), a separate measure could be used to quantify the usefulness of a property by seeing how frequently it occurs 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 2 times under the *hound*_{dog} sub-branch. In contrast, 'large' occurs 36 times under *dog*_{canine} but 10 times under *hound*_{dog}. So *small* is more differential than *large* in the context of hounds. Furthermore 'small' occurs frequently enough to be considered an important attribute for dogs, unlike 'bred by Pharaohs,' which only occurs once (for *Ibizan hound*).

CHAPTER 4 DIFFERENTIA REFINEMENT

After the conceptual differentia properties have been extracted from a definition (as discussed in last chapter), the words for the relation source and object terms should be disambiguated to order to reduce vagueness in the relationships. In addition, the relation types should be refined from surface-level relations or relation-indicating words (e.g., prepositions) into the underlying semantic relationship.¹ Both aspects of this refinement are discussed in this chapter with emphasis on relation refinement because word-sense disambiguation is now relatively mature (Kilgarriff and Palmer, 2000b; Edmonds and Kilgarriff, 2002).

Note that the two refinement processes, namely relation term disambiguation and relation type refinement, 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 intertwine the processes iteratively. This way the results of disambiguation for the some of the source and target terms can influence the refinement of the relation types (and vice versa).

4.1 Source and Target Term Disambiguation

The first step in differentia refinement is to resolve the relational source and target terms into the underlying concepts. 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 closed related, so word-sense disambiguation (WSD) 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 need 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 6.2.1).

For WordNet, the definitions have recently been sense-tagged as part of the Extended WordNet project (Harabagiu et al., 1999; Novischi, 2002). The main approach here just incorporates these sense annotations. For other dictionaries, use of traditional word-sense disambiguation algorithms would be

¹For clarity, *relationships* refers to relation instantiations, and *relations* to the types.

required. Therefore, a few approaches are sketched here for the sake of completeness.

4.1.1 Word-sense Disambiguation of Dictionary Definitions

Three distinct approaches are presented for disambiguating the terms from the definitions. These differ in the amount of training data that is needed beforehand and 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 they 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.

4.1.1.1 Supervised WSD

The standard approach to statistical WSD is based on example-based learning over word-sense annotations. (For background on example-based learning, see Appendix B.) Figure 4.1 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 4.2 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 turn out to be an important clue for word-sense disambiguation. Collocational features are typically binary and indicate the presence of a word that is strongly associated with a particular sense of the word.

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

⟨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 4.1: **Sample word-sense annotations for 'circuit' from Senseval II.**

| | |
|-------------------------|--|
| Morph: | morphology of the target word |
| POS−i: | part-of-speech of <i>i</i> th word to left |
| POS+i: | part-of-speech of <i>i</i> th word to right |
| Word−i: | <i>i</i> th word to the left |
| Word+i: | <i>i</i> th word to the right |
| WordColl _s : | occurrence of word collocation for sense <i>s</i> in context |

Figure 4.2: **Features for supervised word-sense disambiguation.**

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). Thus, a supervised training approach is not yet feasible for resolving the senses of all of the relational objects.

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 25K 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.

4.1.1.2 Unsupervised WSD

Given the limitations of supervised WSD, unsupervised approaches are more suitable for resolving the specific relational object terms. A simple but effective approach is the definition word-overlap approach developed by Lesk (1986), in which 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. Rosenzweig developed a version of this incorporating TF-IDF weighting for the overlap terms that performed quite well in SENSEVAL I as discussed in (Kilgarrieff and Rosenzweig, 2000).

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 1000+ categories in Roget's Thesaurus. Then 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. See Figure 4.3. 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.

4.1.1.3 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.

1. Train Roget classifier over corpus

N = total number of words

N_{cat} = number of words associated with Roget category cat

$\text{freq}(\text{word})$ = number of times word co-occurs in corpus

$\text{freq}_{\text{cat}}(\text{word})$ = number of times word co-occurs with any word in cat

$\hat{P}(\text{cat}) = 1/\#\text{categories}$

$\hat{P}(\text{word}) = \text{freq}(\text{word})/N$

$\hat{P}(\text{word}|\text{cat}) = \text{freq}_{\text{cat}}(\text{word})/N_{\text{cat}}$

2. Disambiguate words by finding category with most support

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

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

Figure 4.3: **Word-sense disambiguation using Roget-based classifier.** Steps 1 and 2 are based on (Yarowsky, 1992). Step 3 is an extension for using this for WSD using WordNet distinctions.

1. Apply named-entity tagging.
For example, person-name \Rightarrow person#1 (likewise for organizations and locations).
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. Determine overlap of noun-context's 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. Select the sense that has largest overlap provided that next largest is a certain threshold away.
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 same synset.
8. Tag non-disambiguated words at semantic distance of 1 with sense for related synsets.

Figure 4.4: ***Heuristics for semi-supervised WSD using bootstrapping.*** Adapted from (Mihalcea and Moldovan, 2001). This forms the basis for one of the WSD systems used in the preparation of Extended WordNet (Novischi, 2002).

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 4.4 shows the heuristics used by their system.

4.1.2 Using Sense Annotations from Extended WordNet

The Extended WordNet 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).²

Figure 4.5 gives an example of the XML 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):

a small_{adj#1} short_{adj#1}-legged_{adj#1} smooth_{adj#1}-coated_{adj#1} breed_{n#2}
of hound_{n#1}

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 4.4) 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”). In addition, they have also done manual checking for 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 taggings 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 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

²Extended WordNet version 2.0.1-1 is used here. The database is freely available at <http://xwn.hlt.utdallas.edu>.

```

<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>-</wf>
    <wf pos="JJ" lemma="legged" quality="silver" wnsn="1">legged</wf>
    <wf pos="JJ" lemma="smooth" quality="normal" wnsn="1">smooth</wf>
    <punct>-</wf>
    <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 4.5: *Extended WordNet annotations for 'beagle' definition.*

roles that 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 just use a small number of very general roles like *beneficiary*. At the other extreme, some systems use very 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). General-purpose systems have typically employed limited case role inventories, with specialized roles mainly being used by task-oriented systems (Dyer, 1983). Palmer (1990) discussed some of the more contentious issues regarding case systems, including adequacy for representation, such as in reliance solely upon case for 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; 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) is as one example of this, with synsets being defined in terms of other synsets rather than using a set of features like [\pm ANIMATE]. At the same time, there has been a shift in emphasis 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 a 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 corpus. Relation refinement experiments for both of these role inventories are discussed later in this chapter. In addition, the future work chapter discusses how the knowledge can be transferred between the two resources.

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:

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

| Role | Freq. | Description |
|-------------|-------|--|
| locative | .092 | place/setting of the event |
| temporal | .120 | indicates when, how often, or how long |
| direction | .030 | starting or ending location (trajectory) |
| manner | .023 | indicate 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). By way of comparison, the *subject* role occurs 41% of the time.

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 are 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) attribute indicates the semantic roles, and the phrase type (PT) attribute indicates 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 from Knowledge Representation Frameworks

The next two case inventories discussed, from Cyc and Conceptual Graphs respectively, are based 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 refinement 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 for differentia extraction, as discussed in Section 4.3.5.

4.2.3.1 Cyc

The Cyc (Lenat, 1995) system 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, portions of the KB have been made freely available public to encourage broader usage. 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.

Cyc uses a wide range of role types from general roles like *beneficiary* through commonly occurring participation types like *victim* to situation-specific roles like *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.

| Role | Freq. | Description |
|-------------------|-------|---|
| speaker | 8310 | person producing the message |
| message | 7103 | content which is communicated |
| self-mover | 6778 | living being moving under its own power |
| theme | 6403 | object changing location or in some spatial relation to a particular location |
| agent | 5887 | entity that acts on another entity |
| goal | 5560 | identifies the endpoint of movement |
| path | 5422 | trajectory which is neither a source nor a goal |
| cognizer | 4585 | person who becomes aware of a phenomenon |
| manner | 4474 | property of motion unrelated to the trajectory |
| source | 3706 | starting-point of motion |
| content | 3662 | entity whose salience is described |
| experiencer | 3567 | being who has a physical experience |
| evaluatee | 3108 | entity about whom/which a judgment is made |
| judge | 3107 | assumed judge who forms the opinion of protagonist's mental properties |
| topic | 3074 | subject matter of the communicated message |
| undefined | 2531 | undefined frame element |
| cause | 2306 | non-agentive cause of the physical experience |
| addressee | 2266 | person that receives a message |
| perceptual source | 2179 | source of perception (e.g., clatter [of hoofs]) |
| phenomenon | 1969 | entity or phenomenon that the perceiver experiences with his or her senses |
| reason | 1789 | reason for the judgment |
| area | 1328 | a region in which the motion takes place |
| degree | 1320 | no description; ex: shook me [to my roots] |
| body part | 1230 | the body part in which a sensation is located |
| protagonist | 1106 | 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 (FrameDescs.html).

| OpenCyc Thematic Roles | | |
|------------------------------|------|---|
| Role | Freq | Description |
| doneBy | 473 | This predicate relates an event to its "doer". |
| performedBy | 317 | doer deliberately does act. |
| objectOfState-Change | 214 | object undergoes some kind of intrinsic change of state |
| objectActedOn | 151 | object is altered or affected in event |
| outputsCreated | 136 | object comes into existence sometime during event |
| transporter | 118 | object facilitating conveyance of transportees |
| transportees | 118 | object being moved |
| toLocation | 110 | where the moving object is found when event ends |
| objectRemoved | 96 | object removed from its previous location |
| inputs | 95 | pre-existing event participant destroyed or incorporated into a new entity |
| products | 94 | obj is one of the intended outputs of event |
| inputsDestroyed | 92 | object exists before event, is destroyed during event |
| fromLocation | 89 | loc is where some moving-object in the move is found at the beginning |
| primaryObject-Moving | 88 | object is in motion at some point during the MovementEvent move and this movement is focal in move |
| seller | 81 | agent sells something in the exchange |
| objectOf-Possession-Transfer | 81 | rights to use object transferred from one agent to another |
| transferredThing | 80 | obj is being moved, transferred, or exchanged in the GeneralizedTransfer event transfer |
| senderOfInfo | 79 | sender is an agent who is the source of information transferred |
| inputsCommitted | 75 | object exists before event and continues to exist afterwards, and as a result of event, object becomes incorporated into something created during event |
| objectEmitted | 69 | obj is emitted from the emitter during the emission event |

Table 4.3: ***Most common thematic roles in OpenCyc.***

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 discussed later. In addition, the future work chapter sketches out how annotations for attributes can be inferred from Cyc. This extends the relation marker inference technique discussed later 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 in the appendix, such as *destination* and *initiator*. 37 conceptual relations in all 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.

Currently there does not seem to be either a resource containing CG-style annotations or a CG-style KB with the semantic relations. Therefore, the role usage is estimated by doing web searches for the various relation names and abbreviations and seeing when CG-style notation is used with the relation, such as in academic papers. 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

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

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

check for the arrows is omitted, as the linear notation makes the arrows optional in certain contexts. Table 4.4 shows the relative frequencies of the roles 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 $\rightarrow(\text{relation})\rightarrow$ or $\leftarrow(\text{relation})\leftarrow$ constructions can 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.³ 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 semantic network (Cassidy, 2000) is a knowledge base derived initially from the 1911 version of Roget's Thesaurus. Part of purpose is to make explicit the relations that hold between the Roget categories and the words listed in each entry. It incorporates information from other resources as well, notably the Unified Medical Language System (UMLS), which formed the basis for the initial set of semantic relations.

Figure 4.6 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.⁴ There are over 400 relations instantiated in the semantic network. Some of these are quite specialized (e.g., *has-brandname*). In addition, there are quite a few inverse functions, since most of the relations are not symmetrical. Certain features of the semantic network representation are ignored during the relation extraction. For example, relation specifications can have qualifier prefixes, such as an ampersand to indicate that the relationship only sometimes holds.

³Factotum 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.

⁴For clarity, some of the relations are renamed to make the directionality more explicit, following a suggestion for their interpretation in the Factotum documentation.

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.

⇒

| | |
|--|--|
| <change, <i>has_subtype</i>, conversion> | <change, <i>has_subtype</i>, transformation> |
| <conversion, <i>has_case</i>, initial state> | <conversion, <i>has_case</i>, final state> |
| <conversion, <i>has_patient</i>, object> | <conversion, <i>has_patient</i>, entity> |
| <conversion, <i>is-result-of</i>, converting> | <conversion, <i>is-result-of</i>, transforming> |
| <process, <i>has_subtype</i>, converting> | <process, <i>has_subtype</i>, transforming> |
| <conversion, <i>has_subtype</i>, processing> | |

Figure 4.6: **Sample entry from Factotum with extracted relations.**

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 later.⁵ The functional relations are shown in boldface. In particular, the meronym or part-whole relations (e.g., *is-conceptual-part-of*) are not included. This accords with their classification by Cruse (1986) as hierarchical relations. Note that the usage counts 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.

Table 3.2 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 only averages 1.3 non-taxonomic properties per concept (including inverses). Factotum compares favorably in this respect, averaging 1.8 properties per concept. OpenCyc provides the highest average at 3.7 properties per concept, although with an emphasis on argument con-

⁵The database files and documentation for the semantic network are available from Micra, Inc. via <ftp://micra.com/factotum>.

| Relation | Usage | Description |
|--|-------|---|
| has-subtype | 37355 | inverse of <i>is-a</i> relation |
| is-property-of | 7210 | object with given salient character |
| is-caused-by | 3203 | indicates force that is the origin of something |
| has-property | 2625 | salient property of an object |
| has-part | 2055 | a part of a physical object |
| has-high-intensity | 1671 | intensifier for the property or characteristic |
| has-high-level | 1564 | implication for the activity (e.g., intelligence) |
| is-antonym-of | 1525 | generally used for lexical opposition |
| is-conceptual-part-of | 1408 | parts of other entities (in case relations) |
| has-metaphor | 1313 | non-literal reference to the word |
| causes _{mental} | 1208 | motivation (causation in the mental realm) |
| uses | 1157 | a tool needing active manipulation |
| is-performed-by | 1081 | human actor for the event |
| performs _{human} | 987 | human role in performing some activity |
| is-function-of | 983 | artifact that passively performs the function |
| has-result | 977 | more specific type of <i>causes</i> |
| has-conceptual-part | 937 | generalization of <i>has-part</i> |
| is-used-in | 930 | activity or some desired effect for the entity |
| is-part-of | 898 | distinguishes part from group membership |
| causes | 866 | inverse of <i>is-caused-by</i> |
| has-method | 830 | method used to achieve some goal |
| is-caused-by _{mental} | 810 | inverse of <i>causes</i> _{mental} |
| has-consequence | 785 | causation due to a natural association |
| has-commencement | 663 | state that commences with the action |
| is-location-of | 655 | absolute location of an object |
| requires | 341 | object or sub-action necessary for an action |
| is-studied-in | 331 | inquires into any field of study |
| is-topic-of | 177 | communication dealing with given subject |
| produces | 166 | what an action yields, generates, etc. |
| is-measured-by | 158 | instrument for measuring something |
| is-job-of | 117 | occupation title for a job function |
| is-patient-of | 101 | action that the object participates in |
| is-facilitated-by | 98 | object or sub-action aiding an action |
| is-biofunction-of | 27 | biological function of parts of living things |
| was-performed-by | 22 | <i>is-performed-by</i> occurring in the past |
| has-consequence _{object} | 21 | consequence for the patient of an action |
| is-facilitated-by _{mental} | 9 | trait that facilitates some human action |

Table 4.5: **Sample relations from Micra's Factotum.** Boldface relations are used in the experiments discussed later in Section 4.3.4.2.

straints and other usage restrictions.⁶ Therefore, the information in Factotum complements WordNet through the inclusion of more functional relations.

4.3 Relation Refinement

The goal of relation refinement 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 refinement can be seen as a special case of word-sense disambiguation. As an example, refining the relationship ⟨‘dog’, ‘with’, ‘ears’⟩ into ⟨‘dog’, *has-part*, ‘ears’⟩, is equivalent to disambiguating the preposition ‘with,’ given that the senses are 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 refinement of implicitly indicated relations is an area for future work.

Traditionally, prepositions have numerous senses. WordNet does not include function words, so sense inventories from other dictionaries are discussed here. For instance, the preposition ‘for’ has 20 different senses defined in Merriam-Webster’s dictionary (10th Edition),⁷ as shown in Table 4.6. Since the Treebank roles are more general than these, the disambiguation in the first set of experiments address a coarse form of sense distinction. For the preposition ‘for,’ there are six distinctions (or four with low-frequency pruning). In contrast, since the FrameNet distinctions are quite specific, the disambiguation in the second set of experiments address fine-grained sense distinctions. For ‘for,’ there are 41 distinctions (or 18 with low-frequency pruning).

4.3.1 Overview of Relation Type Disambiguation

Unlike the situation with relational source and object terms, there is just a limited number of relation types. For example, CYC has the largest number of thematic roles compared to other resources, but the total number is still just a few hundred (Lehmann, 1996), which is quite small compared to the 100,000+

⁶These figures are derived by counting the number of relations excluding the instance and subset ones. Cyc’s comments and lexicalizations are also excluded (implicit in Factotum and WordNet). The count is then divided by the number of concepts.

⁷An online version of Merriam-Webster’s is available at www.m-w.com.

| Sense | Definition |
|-------|--|
| 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, to vote for a levy] |
| 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, to fight for one's life] |
| 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, an ear for music] |
| 12. | as being [to know for a fact] |
| 13. | considering the nature of; as concerns [cool for July, clever for a child] |
| 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: ***Definition of preposition 'for' in Merriam-Webster's dictionary.***

synsets in WordNet. Therefore, a supervised learning approach is much more feasible. In this approach, predefined classifications for several examples of each of the different relation types (or *roles*)⁸ are input into a machine learning system that learns classification rules (either symbolic or statistical).

4.3.1.1 Use of Class-based Collocations

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 (e.g., general-purposes newspaper text rather than 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 the words in a window around the target word and 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 an amount greater than 20%. That is, the relative difference between the conditional and prior probabilities for the class value must 20% or higher for the word to be treated as one of its collocation:

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

The second pass over the training data determines the value for the collocational feature of each classification category by checking whether the current

⁸The 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 (referred to as ‘relationship’ here).

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. 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-independence 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 4.7. 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 that occur within 5 words of the target prepositions are considered.⁹

The main difference from that of a standard WSD approach is that, during the determination of the class-based collocations, each word token is replaced by synset tokens for its hypernyms in WordNet, several of which might occur more than once. This introduces noise due to ambiguity, but given the conditional-independence selection scheme, the preference for hypernym synsets that occur for different words compensates. The feature settings in Figure 4.7 are used in two different configurations: word-based collocations alone, and a combination of word-based and hypernym-based collocations. The combination generally produces the best results. This combines the specific clues provided by the word collocations with improved generalizations provided by the hypernym collocations.

4.3.2 Penn Treebank

When deriving training data from Treebank from the parse tree annotations, the functional tags associated with prepositional phrases are converted into preposition sense tags. Consider the sample annotation for Treebank shown earlier:

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

Treating this as the preposition sense would yield the following annotation:

This every man contains within_{LOC} him.

Frequency counts for the prepositional phrase case role annotations are shown in Table 4.7.

⁹This 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.

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_i: word collocation for role *i*
HypernymColl_i: hypernym collocation for role *i*

Collocation Context:

Word: anywhere in the sentence
Hypernym: within 5 words of target preposition

Collocation selection:

Frequency: $f(\text{word}) > 1$
Conditional probability threshold: $P(C|\text{coll}) \geq 0.5$
Conditional independence threshold: $(P(C|\text{coll}) - P(C))/P(C) \geq .20$
Organization: per-class-binary

Model selection:

overall classifier: Decision tree
individual classifiers: Naive Bayes
10-fold cross-validation

Figure 4.7: **Feature settings used in preposition classification experiments.** The *per-class-binary* organization uses a separate binary feature per role (Wiebe et al., 1998a).

| Tag | Role | Freq. |
|-----|-------------|-------|
| LOC | locative | 17220 |
| TMP | temporal | 10572 |
| DIR | direction | 5453 |
| MNR | manner | 1811 |
| PRP | purpose | 1096 |
| EXT | extent | 280 |
| BNF | benefactive | 44 |

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* is frequency of the role occurrences.

| Relation | P(R) | Example |
|-----------|------|--|
| locative | .732 | workers <i>at</i> a factory |
| temporal | .239 | expired <i>at</i> midnight Tuesday |
| manner | .020 | has grown <i>at</i> a sluggish pace |
| direction | .006 | CDs aimed <i>at</i> 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.

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 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,

| 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. $P(R|C)$ is probability of the relation given that the synset category occurs in the context.

$$\begin{aligned}
P(R = \text{locative} | C = \text{entity\#1}) - P(R = \text{locative}) &= .13 \\
P(R = \text{temporal} | C = \text{entity\#1}) - P(R = \text{temporal}) &= -.12 \\
P(R = \text{other} | C = \text{entity\#1}) - P(R = \text{other}) &= -.22,
\end{aligned}$$

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,

$$\begin{aligned}
P(R = \text{locative} | C = \text{abstraction\#6}) - P(R = \text{locative}) &= -.22 \\
P(R = \text{temporal} | C = \text{abstraction\#6}) - P(R = \text{temporal}) &= .22 \\
P(R = \text{other} | C = \text{abstraction\#6}) - P(R = \text{other}) &= .001,
\end{aligned}$$

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 both types of 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

| Experiment | Accuracy | STDEV | # Instances: 26616 |
|------------|----------|-------|--------------------|
| Word Only | 81.1 | .996 | # Classes: 7 |
| Hypernym | 85.9 | .702 | Entropy: 1.917 |
| Combined | 86.1 | .491 | Baseline: 0.480 |

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* selects the most-frequent role. The *Word Only* experiment just uses 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 trails. The difference in the two experiments is statistically significant at $p < .01$.

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: 78.5% versus 77.8% accuracy.

4.3.3 FrameNet

A similar preposition word-sense disambiguation experiment is carried out over the FrameNet semantic role annotations that apply to 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 overPath
its back.

```

| 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.

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

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

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.

4.3.3.1 Illustration with 'at'

It is illustrative to compare the prior probabilities (i.e., P(R)) for FrameNet to those seen earlier for 'at' in Treebank. See Table 4.12 for the most frequent

| 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.**

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 from Treebank is not directly represented in FrameNet, and it is not subsumed by another specific role. Also, while there is a *location* role in FrameNet, it applied less in 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 not generally annotated.

4.3.3.2 Results

Table 4.14 shows the results of classification when all of the prepositions are classified together. The overall results are not that high due to the very large number of roles. 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 less than 1% of the role occurrences for prepositional phrases, significant improvement results, as shown in Table 4.15. Even with this filtering, the classification is still 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 prone to being misclassified with roles that they might not have occurred with in the training data, such as whenever other contextual

| Experiment | Accuracy | STDEV | # Instances: | 27295 |
|------------|----------|-------|--------------|-------|
| Word Only | 48.9 | 0.94 | # Classes: | 129 |
| Hypernym | 48.0 | 1.32 | Entropy: | 5.128 |
| Combined | 49.4 | 0.59 | Baseline: | 0.149 |

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

| Experiment | Accuracy | STDEV | # Instances: | 22125 |
|------------|----------|-------|--------------|-------|
| Word Only | 59.5 | 1.20 | # Classes: | 25 |
| Hypernym | 58.4 | 1.32 | Entropy: | 4.055 |
| Combined | 60.5 | 1.14 | Baseline: | 0.184 |

Table 4.15: **Preposition disambiguation without low-frequency FrameNet roles.** Relations that occur less than 1% of the total are excluded. See Table 4.10 for the legend.

clues are strong for that role. This is not a problem with Treebank given its small role inventory.

4.3.4 Factotum

Note that FACTOTUM does not indicate the way the relationships are expressed in English. WordNet similarly does not indicate this, but does include definition glosses that can be used in some cases to infer the *relation markers* (i.e., generalized case markers). For example,

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

WordNet: *dry*_{alter} remove the moisture from and make dry
*dryer*_{appliance} an appliance that removes moisture

Therefore, the Factotum relations cannot be used as is 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.

| 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-word results for preposition disambiguation with FrameNet.*** See Table 4.11 for the legend.

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 search¹⁰, 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 prepositions¹¹ are included in a series of proximity searches to find the preposition occurring the most with the terms. For instance, given the relationship \langle drying, *is-function-of*, drier \rangle , the following searches would be performed.

drying NEAR drier NEAR of
drying NEAR drier NEAR to
...
drying NEAR drier NEAR "because of"

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:

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

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 \langle drying, *is-function-of*, drier \rangle relationship are 'during,' 'after,' and 'with.'

This technique can readily be extended to finding relation markers in foreign languages, such as Spanish, given a bilingual dictionary. Ambiguous translations pose a complication, but in most of these cases, similar relation markers should be likely unless the relations between the alternative meaning pairs diverge significantly.¹² For example, when the process is applied

¹⁰AltaVista's Boolean search is available at www.altavista.com/sites/search/adv.

¹¹The common prepositions are determined from the prepositional phrases assigned functional annotations in Penn Treebank II (Marcus et al., 1994).

¹²Sidorov et al. (1999) illustrate the differences that might arise for terms referring to non-adults in English, Spanish, and Russian.

to the translated relationship for the example, namely $\langle \text{secar, is-function-of, secarador} \rangle$, the top three markers are ‘con’, ‘de’, and ‘para.’

4.3.4.2 Method for Classifying the 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 becomes preposition disambiguation, using the Factotum relations as senses.

Figure 4.8 gives the feature settings used in the experiments. This are a streamlined version of the feature set from Figure 4.7, which is used in the Treebank and FrameNet experiments, to account for the lack of sentential context. Figure 4.9 contains sample feature specifications from the experiments discussed in the next section. This shows that ‘n/a’ is used whenever a preposition marker for a particular relationship cannot be inferred. For brevity, the feature specification only includes collocation features for the most frequent relations. Sample collocations are also shown for the relations. 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-independence 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.

4.3.4.3 Results

For this task, the set of functional relations in Factotum 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.

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_i: 1 iff context contains any word collocation for relation i
 HypernymColl_i: 1 iff context contains any hypernym collocation for relation i

Collocation selection:

- Frequency constraint: $f(\text{word}) > 1$
 Conditional independence threshold: $(P(C|\text{coll}) - P(C))/P(C) \geq .20$
 Organization: per-class-binary grouping

Model selection:

- Decision tree using Weka's J4.8 classifier (Witten and Frank, 1999)
 10-fold cross-validation

Figure 4.8: **Features used in Factotum role classification experiments.**

| <i>Experiment</i> | <i>Accuracy</i> | <i>Stdev</i> | # Instances: 5959 |
|-------------------|-----------------|--------------|-------------------|
| Word | 68.4 | 1.28 | # Classes: 21 |
| Hypernym | 53.9 | 1.66 | Entropy: 3.504 |
| Combined | 71.2 | 1.78 | 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 $p < 0.01$ (via a paired t-test).

The dataset also includes the inferred relation markers, thus introducing some noise.

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. Note that the classification task is quite challenging, given the large number of choices and high entropy (Kilgarriff and Rosenzweig, 2000).

Relationships from Factotum with inferred markers:

⟨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 |
|------------------------|------------------|------------------|---------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| <i>is-caused-by</i> | NN | VB | n/a | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>is-caused-by</i> | NN | NN | by | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>has-consequence</i> | NN | JJ | because | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>has-consequence</i> | JJ | VCN | 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 |
|------------------------|------------------|------------------|---------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| <i>is-caused-by</i> | NN | VB | n/a | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>is-caused-by</i> | NN | NN | by | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>has-consequence</i> | NN | JJ | because | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>has-consequence</i> | JJ | VCN | 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, WC₁, ... WC₇, HC₁, ... HC₇⟩.

where POS_s and POS_t are the parts of speech for the source and target terms, and the relations for the word and hypernym collocations (WC_i and HC_i) follow:

1. *is-caused-by*
2. *is-function-of*
3. *has-consequence*
4. *has-result*
5. *is-caused-by*_{mental}
6. *is-performed-by*
7. *uses*

Figure 4.9: **Sample feature specifications for Factotum experiments.** The collocation features are not shown for the low-frequency relations.

4.3.5 Combining the Different Semantic Role Inventories

For the application to differentia refinement, 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 refinement 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):

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

Most of these are role are contained in the common relation inventory developed for differentia refinement. 26 total relations are defined, including a few roles based on the Treebank, Cyc and Conceptual Graphs inventories. Table 4.18 shows this role inventory along with a description of each case. In addition to traditional thematic relations, this includes a few specialization-type relations. These 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

| 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 <i>Cyc products</i>) |
| Recipient | recipient of the resources |
| Resource | entity utilized during event (same as <i>Cyc inputs</i>) |
| Source | initial position of the theme |
| Theme | entity somehow affected by the event |

Table 4.18: **Common inventory of semantic relations for differentia refinement.** The inventory includes roles primarily based on FrameNet (Fillmore et al., 2001) and Conceptual Graphs (Sowa, 1999); it also includes roles based on the Treebank and Cyc inventories.

| Experiment | Accuracy | STDEV | # Instances: 27295 |
|------------|----------|-------|--------------------|
| Word Only | 54.5 | 0.94 | # Classes: 31 |
| Hypernym | 53.0 | 0.75 | Entropy: 4.006 |
| Combined | 55.5 | 0.54 | Baseline: 0.150 |

Table 4.19: **Overall results for preposition disambiguation with common roles.** The FrameNet annotations are mapped into the common inventory from Table 4.18. See Table 4.10 for the legend.

the 26 common relations shown in Table 4.18. A data file providing this mapping can be found at the following URL:

www.cs.nmsu.edu/~tomohara/differentia-extraction/relation-mapping.html.

This data file also specifies mappings for the other inventories discussed in this chapter, such as from OpenCyc into the common inventory. Results for the classification of the FrameNet data mapped into the common inventory are shown in Table 4.19. 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 to 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 refinement. 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 differentia.

4.4 Differentia Refinement Algorithm

To summarize the approach taken to differentia refinement, Figure 4.10 presents the high-level algorithm for the process. Differentia refinement is done in two main steps. First, the source and target terms are disambiguated. Next, the relation-indicating terms are refined into semantic relations by applying the

common-inventory relation classifier. Support for refining relations indicated by 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.

The next chapter shows how this refinement process facilitates lexical augmentation, using a combination of the semantic role data from Treebank and FrameNet. Note that the integration of the data from Factotum is not addressed due to time constraints.

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 “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)

⇒

(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 4.1.

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 4.10: ***Differentia refinement 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 refined 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). Evaluation is based on having humans assess the quality of random samples from the extracted relationship listings. The second area is word-sense disambiguation. Evaluation is based on comparing the performance of systems utilizing the differentiating relationships (i.e., differentia) versus those based on standard approaches.

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 refinement 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 3.1. Of these, 10,810 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 much 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 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

| 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 | Experience | 0.017 |
| ACCM | Accompaniment | 0.011 |
| AREA | Area | 0.010 |
| DIR | Direction | 0.001 |

Table 5.1: **Frequency of relations after refinement.** WordNet definitions are analyzed with relations refined into the common relation inventory (Table 4.18).

data for the annotations (see Table 4.18). 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 like *modification* that were not considered during the refinement process. Therefore, no extracted relations are listed for the last case (*verb:lunge#1*).

5.1.2 Evaluation

Four human judges were recruited to evaluate random samples of the relations that were extracted. All are graduate students in computer science with exposure to computational linguistics, and each was given 100 relationships to evaluate. To allow for inter-coder reliability analysis, each evaluator evaluated 50 samples that were also evaluated by the others, half as part of a training phase and half after training. In addition, they also evaluated 20 samples that

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 4.10. Syntactic relations such as modification are omitted for sake of brevity. Relation strengths derived via cue validity analysis are shown in parentheses.

were manually corrected beforehand. This provides a baseline against which the uncorrected results can be measured against.

Because the thesis only addresses relations indicated by prepositional phrases, the evaluation is restricted to these cases. The evaluation also does not directly address aspects related to prepositional attachment, although incorrect attachment decisions by the parser does negatively effect the evaluation. Specifically, the judges rate the assignment of relations to the prepositional phrases on a scale from 1 to 5, with 5 being an exact match. They are presented with the list of common relation types shown in the last chapter (Table 4.18), so that correctness is relative to this relation inventory.

For example, consider the ‘kennel’ example from the last chapter:

A kennel is an outbuilding that serves as a shelter for a dog.

\langle verb:serves#1, *Reason*, noun:dog#1 \rangle
 \langle verb:serves#1, *Manner*, noun:shelter#1 \rangle

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; but, in general, more detailed instructions are desirable to achieve better uniformity.

5.1.2.1 Inter-coder Reliability Analysis

To assess the reliability of the evaluations, the *Kappa Statistic* was calculated:

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

This determines the extent to which the coders agree (P_a) less that which is due to chance agreement (P_e). Although the formula for kappa itself is simple, the derivation of the intermediate values is somewhat complicated, as shown in Figure 5.3.

The frequency of each of the categories (e.g., 1 through 5) is tabulated across all judgments. Actual agreement is based on averaging for each distinct coding event, the percentage of the possible number of pairwise agreements.

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, off#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 4.18.

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, off#GROUND, noun:cause> (bad) 1 __2 _3 _4 x_5 _ (good)`

`<noun:cause, off#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 4.18).

Let $M = \text{\#events}$, $N = \text{\#categories}$, and $K = \text{\#coders}$:

| | | |
|-------------|---|--|
| n_{ij} | judgments for category j in event i | |
| A_i | agreement for event i | $\sum_{j=1}^N \frac{n_{ij}(n_{ij}-1)/2}{K(K-1)/2}$ |
| P_a | actual agreement | $\frac{\sum A_i}{M}$ |
| C_j | frequency for category j | $\sum_{i=1}^M n_{ij}$ |
| \hat{p}_j | estimated probability of category j | $\frac{C_j}{MK}$ |
| P_e | expected agreement | $\sum_{j=1}^N \hat{p}_j \times \hat{p}_j$ |

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

Figure 5.3: **Calculation of the Kappa statistic.** Based on (Siegel and Castellan, 1988; Carletta, 1996).

Note that given K coders, the maximal number of such pairings is K choose 2. As shown in the figure, n_{ij} is the number of judgments for a particular component value. So n_{ij} choose 2 is the number of pairwise agreements for that component value. Therefore, the formula for A_i just adds up the number of all pairwise agreements for that event and normalizes it by the maximal number. Then the actual agreement (P_a) is calculated as the average of all the A_i 's. The frequency of judgments for each category is tabulated across all events. The probability of each category is then estimated by dividing the category frequency of the total number of judgments. Then the expected agreement (P_e) is defined as the weighted average of these probabilities. In other words, it is the expected value for the probability of a generic category. Lastly, the kappa value is calculated as the ratio of the agreement not due to chance over the corresponding maximum value. Note that kappa can thus be negative if the actual agreement is less than that due to chance.

A κ value of 0.8 or greater indicates a high level of reliability among raters, with values between 0.67 and 0.8 indicating only moderate agreement (Carletta, 1996). Table 5.2 shows the results of the inter-coder reliability analysis over the judge rating given to the manually corrected subset of the samples evaluated. The overall kappa score is quite low (.143), but this is partly because

| Categories | Instances | P_a | P_e | κ |
|------------|-----------|-------|-------|----------|
| 5 | 25 | .333 | .222 | .143 |
| 3 | 25 | .647 | .426 | .384 |
| 2 | 25 | .740 | .516 | .463 |

Table 5.2: **Inter-coder reliability analysis for evaluation.** 25 relationships were each evaluated by 4 judges. *Categories* gives the number of ranges for the assessment scores: with 3 categories, scores 1&2 and 4&5 are treated the same; and, with 2 categories, scores 1&2 and 3-5 are treated the same. *Instances* is the number of distinct assessments per coder. P_a gives the actual agreement, P_e gives the expected agreement, and κ gives the kappa statistic.

the agreement is based on exact matching of assessment scores. 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 kappa measure increases to .384. The scores are also converted into 2 values by combining 1 and 2 as *bad* as well as 3 through 5 as *ok*, resulting in a kappa measure of .463.

NOTE: These agreement results are just for the dry-run evaluation. I expect better agreement results for the actual evaluation, as a result of feedback on coding errors and more clarification of the relations.

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 3.225, which translates into an overall score of 0.60. For the uncorrected system output, the mean assessment value was 3.033, which translates into an overall score of 0.58. 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 the standard supervised approach for WSD using tagged training data to induce a statistical classifier for each word to be disambiguated.

| Corrected | | Uncorrected | |
|-----------|-------|-------------|-------|
| Cases | 10 | Cases | 15 |
| Scores | 40 | Scores | 60 |
| Mean | 3.225 | Mean | 3.033 |
| STDEV | 1.625 | STDEV | 1.551 |
| Score | 0.60 | Score | 0.58 |

Table 5.3: **Mean assessment score for all extracted relationships.** 25 relationships were each evaluated by 4 judges (same as with Table 5.2). *Corrected* shows assessments over manually-corrected output, whereas *Uncorrected* evaluates the system output as is. *Cases* is the number of distinct relationships, and *Scores* is the number of individual assessments. *Mean* gives the mean of the assessment ratings (from 1 to 5), and *STDEV* is the corresponding standard deviation. *Score* is score relative to scale from 0 to 1.

The second is a hybrid supervised/unsupervised approach that incorporates knowledge from WordNet to provide a model of the dependencies among word senses.

5.2.1 Supervised Classification

The standard approach using statistical classification for word-sense disambiguation was illustrated in Section 4.1.1.1 of the previous chapter. The same general approach is used, 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 +1). The POS_{+0} feature is labeled *Morph* in Figure 4.2, 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). The collocation variable $WordColl_i$ for each sense S_i is binary, corresponding to the absence or presence of any word in a set specifically chosen for S_i . A word W is considered as

Features:*Standard features*

| | |
|-------------------------|--|
| Morph: | morphology of the target word (i.e., part of speech) |
| POS _{-i} : | part-of-speech of <i>i</i> th word to left |
| POS _{+i} : | part-of-speech of <i>i</i> th word to right |
| Word _{-i} : | <i>i</i> th word to the left |
| Word _{+i} : | <i>i</i> th word to the right |
| WordColl _i : | occurrence of word collocation for sense <i>i</i> in context |

Differentia-based features

RelatedColl_i: occurrence of related-word collocation for sense *i* in context

Collocation Context:

Word: anywhere in the sentence

Collocation selection:

| | |
|--|---|
| Word collocation frequency constraint: | $f(\text{word}) > 1$ |
| Related-Word collocation frequency constraint: | $f(\text{related-word}) > 4$ |
| Conditional probability threshold: | $p(c \text{coll}) \geq .50$ |
| Conditional independence (CI) threshold: | $(p(c \text{coll}) - p(c))/p(c) \geq .20$ |
| Related-Word CI threshold: | $(p(c \text{coll}) - p(c))/p(c) \geq .80$ |
| Feature organization: | per-class-binary (Wiebe et al., 1998a) |

Model selection:

Decision tree via Weka's J4.8 classifier (Witten and Frank, 1999)
10-fold cross-validation

Figure 5.4: **Features for word-sense disambiguation with differentia.** The *RelatedColl_i* features are the only difference from the settings in Figure 4.2.

a collocation for sense S_i if the relative percent gain in the conditional probability over the prior probability is 20% or higher:

$$\frac{(P(S_i|W) - P(S_i))}{P(S_i)} \geq 0.20.$$

However, if the word only occurs once in the training data, it is ignored. This is a variation of the *per-class, binary organization* with a *conditional independence* test used by Wiebe et al. (1998a) There are also four adjacency-based collocational features (*Word $\pm i$* for *i* from -2 to $+1$), which were found to be beneficial in other work (Pedersen and Bruce, 1998; Ng and Lee, 1996). O'Hara et al. (2000) used these feature settings in the first Senseval competition with good results.

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., $P(S_i|W) > .5$). 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 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, in order to evaluate the usefulness of the various types. Here, the purpose is just to show that information derived from differentia analysis is useful for supervised word-sense disambiguation.

The context words are not disambiguated, so the relations for separate senses of the same word are conflated. When determining the potential collocations, 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 independence 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. Also, the words that 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. Also, 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 order 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*) 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 marginal improvements. Overall, the differentia-based system achieves 63.8% accuracy versus 63.4% accuracy for the standard system, with a baseline accuracy of 57.7%.

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 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

Dictionary differentia convey properties that involve lexical relations of varying strengths. For example, 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

| Noun | Senses | Freq | Entropy | Baseline | Standard | System |
|-----------|--------|------|---------|----------|----------|--------|
| 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 using 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; *Standard* uses all the features from Figure 5.4, except for the relatedness collocations derived from differentia (*RelatedColl*); and, *System* includes the relatedness collocations as well.

| Verb | Senses | Freq | Entropy | Baseline | Standard | System |
|-------------|--------|------|---------|----------|----------|--------|
| 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 using Senseval II verb training data.**
See Table 5.4 for the legend.

| Adjective | Senses | Freq | Entropy | Baseline | Standard | System |
|-----------|--------|------|---------|----------|----------|--------|
| 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 using Senseval II adjective training data.*** See Table 5.4 for the legend.

The definition for 'key' indicates a strong relationship to *lock*; but, the definition for 'lock' only indicates a moderate relationship to *door*. Therefore, a probabilistic representation is useful for the representing the differentia. Moreover, given the asymmetries in the relationships, a *Bayesian Network* (Russell and Norvig, 1995; Charniak, 1992; 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 A 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 is on 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 discussed 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* (CV's), 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, the most informative ancestor is estimated based on frequency considerations using Semcor (Miller et al., 1994).

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 that 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, then high-frequency but informative relations like *is-a* would be penalized. Therefore, the relation type and target term are used together as the feature , as in

$\langle *, is-a, mammal\#1 \rangle$.

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 concepts. Therefore, relationships using it are generally weighted high unless the target term is commonly used (e.g., *city#1*). In contrast, relations like *has-attribute* that have a restricted domain (i.e., just attributes and characteristics) are generally weighted low unless the target term is rarely used in such relations (e.g., an uncommon attribute). One aspect that this approach would not account for is the informativeness of relations. For example, a generic semantic-relatedness relation like *related-to* has a large range of target terms and thus is weighted as high as *is-a*, even though it is much less informative. This is left for future work.

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 are 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:

$$\text{output} = \neg((\neg v_1) \wedge \dots \wedge (\neg v_n))$$

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:

$$\begin{aligned} \text{child} &= \neg((\neg v_1) \wedge \dots \wedge (\neg v_n)) \\ P(\text{child} | V_1 = v_1, \dots, V_n = v_n) &= 1.0 - \prod (1.0 - P(\text{child} | v_i)), \quad \forall v_i v_i = T. \end{aligned}$$

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.

As an illustration of the network structure that are derived from lexical relations, consider the definitions for a few types of hounds shown in Figure 5.5. After parsing and then refinement, the resulting semantic relations would be those shown in Figure 5.6. The Bayesian network representation for these relationships would have the same structure as in the figure. To complete the Bayesian network specification, the CPT tables have to be defined. 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 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., $P(N | \dots P_i = \text{True} \dots)$). More details on this derivation process are given in the next section.

As can be seen from the figure, embedded attributes are treated as separate nodes. Therefore, [*beagle* \rightarrow_{has} *coat* $\rightarrow_{\text{attribute}}$ *smooth*] is modeled as [*beagle* \rightarrow_{has} *beagle_coat*] and [*beagle_coat* $\rightarrow_{\text{surface}}$ *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.

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

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

| | |
|-----------------------|-----------------------|
| hound: | wolfhound: |
| is-a dog (1) | is-a hound (0.25) |
| has-part ears (1) | size large (0.333) |
| size large (0.333) | has-part coat (0.5) |
| attr drooping (1) | attr rough (1) |
| used-for hunting (1) | greyhound: |
| basset: | is-a hound (0.25) |
| is-a hound (0.25) | size large (0.333) |
| size small (0.25) | girth slender (1) |
| has-part legs (1) | used-as dog (1) |
| size short (1) | attr racing (1) |
| beagle: | whippet: |
| is-a hound (0.25) | is-a greyhound (0.5) |
| size small (0.25) | size small (0.25) |
| has-part coat (0.5) | location England (1) |
| attr smooth (1) | Italian_greyhound: |
| | is-a greyhound (0.5) |
| | size small (0.25) |
| | attr very (1) |

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

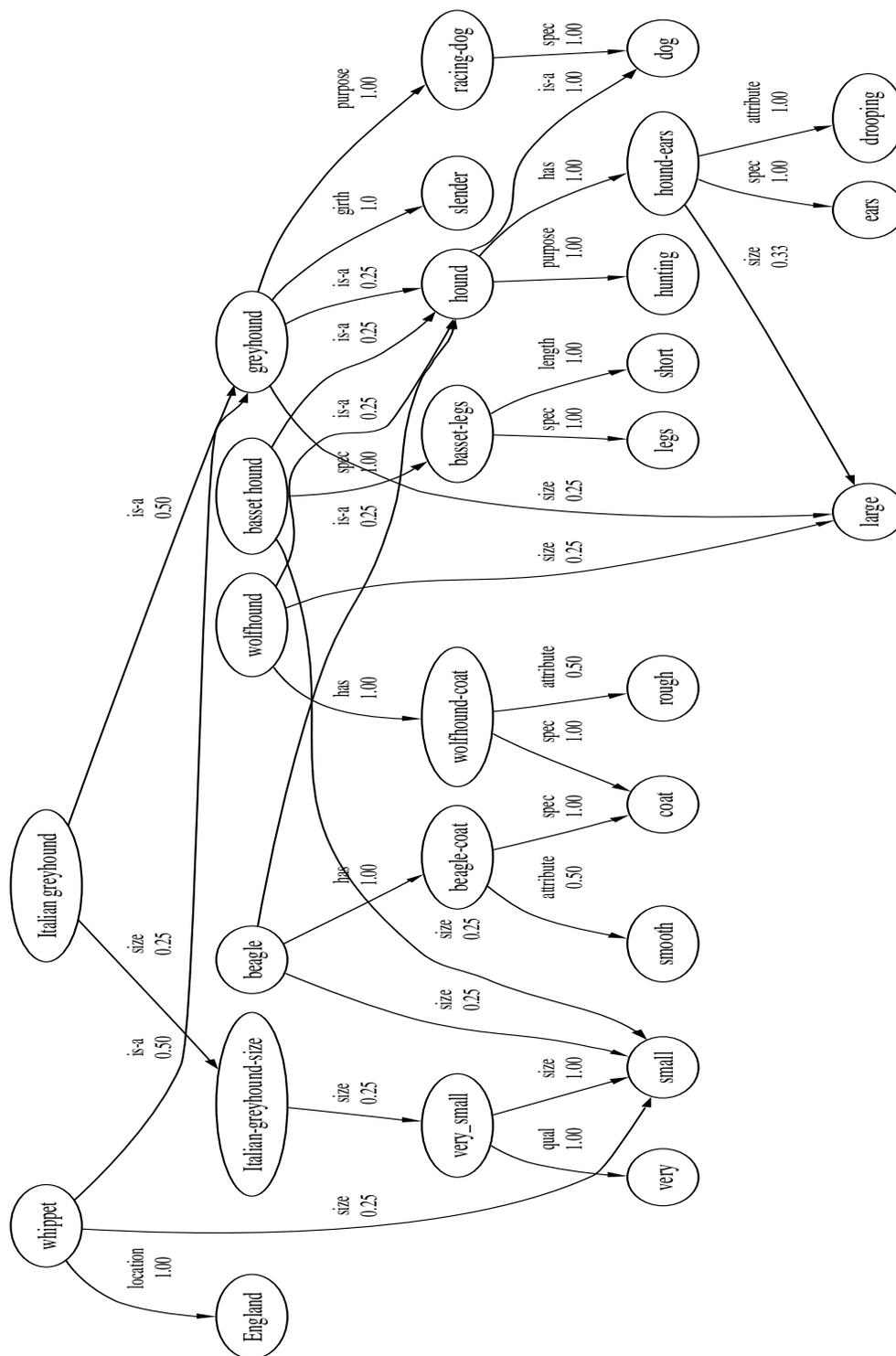


Figure 5.6: **Semantic network of sample hound lexical relations.** The relations shown in Figure 5.5 were used to create semantic network which defines the structure for the Bayesian network.

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.

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. agreement as to goals
6. the body of people in a learned occupation
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 the CPT incorporating the hypernym relations from WordNet, the conditional probability $P(\text{hyponym}|\text{hyponym})$ 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*.

$$P(\text{town\#1} \mid \text{municipality\#1})$$

| municipality#1 | P(town#1) |
|----------------|------------------|
| F | 0.0 + ϵ |
| T | 0.500 |

This illustrates that logical zeros are encoded using an epsilon rather than 0.0. This is a requirement for Bayesian 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 model is used. The following 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:

$$P(\text{municipality\#1} \mid \text{administration_dist\#1}) = 0.0625$$

$$P(\text{municipality\#1} \mid \text{urban_area\#1}) = 0.143$$

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.

$$P(\text{municipality\#1} \mid \text{urban_area\#1}, \text{administration_dist\#1})$$

| urban_area#1 | administration_district#1 | P(municipality#1) |
|--------------|---------------------------|-------------------|
| F | F | 0.0 + 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, dotted links show the semantic-relatedness relations derived from dictionary differentia. 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 effect changes 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*):

$$P(\text{support_town} | \text{town_3}, \text{town_1}, \text{town_2})$$

| town_3 | town_1 | town_2 | support_town |
|--------|--------|--------|--------------|
| F | F | F | € |
| F | F | T | .705 |
| F | T | F | .034 |
| F | T | T | € |
| T | F | F | .264 |
| T | F | T | € |
| T | T | F | € |
| T | T | T | € |

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. The top 100 sentences were chosen for the evaluation. Of these, six 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 baseline system based on (Wiebe et al., 1998b). As can be seen, using a

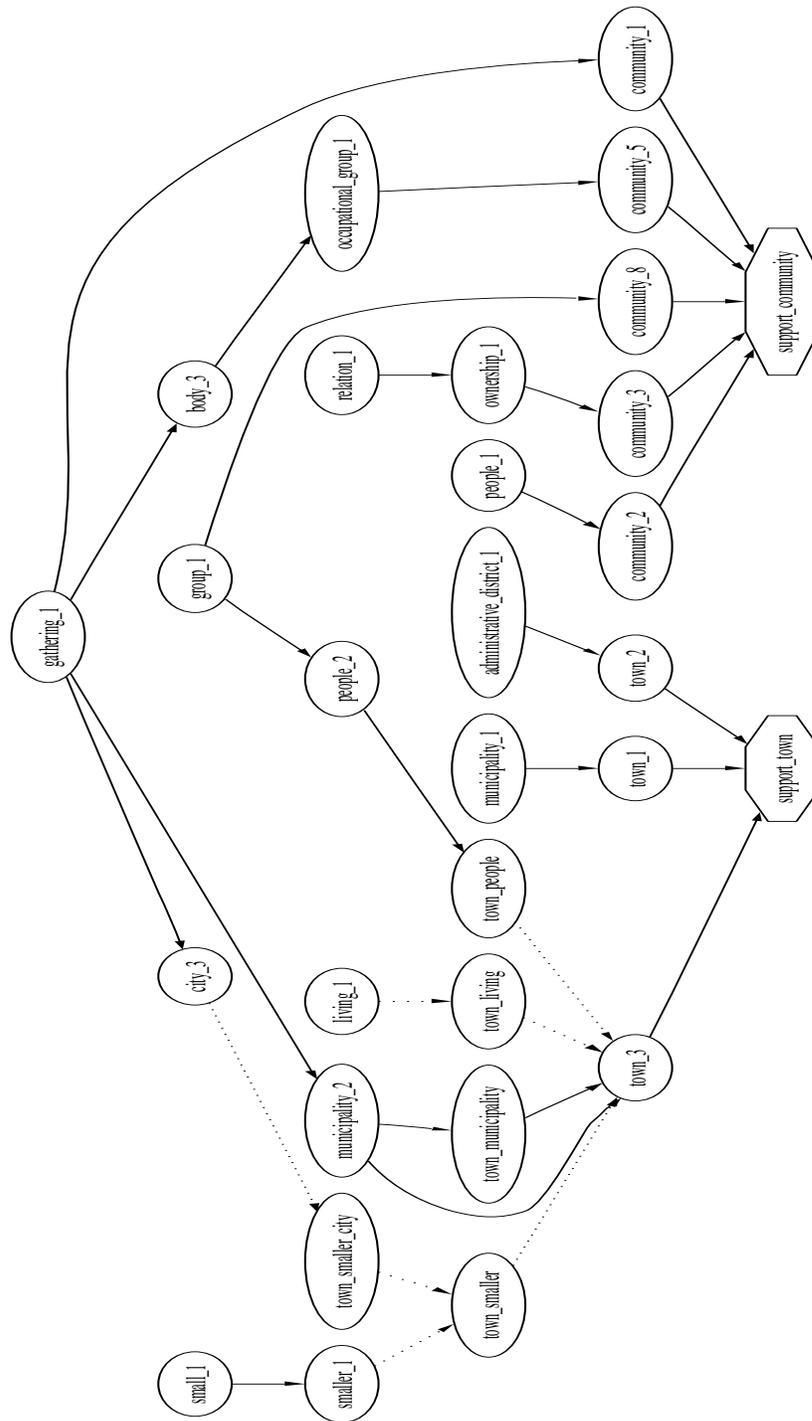


Figure 5.7: **Relations for 'community' and 'town' in WordNet.** Solid links indicate *is-a* relations, and dashed links indicate differentia-based relations. Senses 4, 6, of 7 'community' omitted since not connected to main component.

Bayesian network with differentia-based relations generally leads to improvement over a system that just incorporates the WordNet *is-a* relations. Overall, the differentia-based approach achieves a gain of nearly a two point percentage gain (59.9% versus 61.8%), which is a statistically significant difference.

| SentID | <i>is-a</i> | Both | SentID | <i>is-a</i> | Both | SentID | <i>is-a</i> | Both |
|-----------|-------------|--------|-----------|-------------|--------|-----------|-------------|--------|
| 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 | | | |
| 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* is file and line number for sentence from Wall Street Journal portion of DSO corpus. *is-a* is accuracy of system just using WordNet *hypernym* relations. *Both* is accuracy of system using relations extracted from the definitions glosses in addition to the *is-a* relations. *Total* gives the mean accuracy for each system. The gain in performance is statistically significant at $p < .05$, using a paired t-test.

CHAPTER 6 DISCUSSION AND FUTURE WORK

6.1 Related Work

This work has touched on a variety of areas in computational linguistics. This section discusses the main differences in the approaches taken compared to earlier work.

6.1.1 Differentia Extraction

As seen from the background on lexical semantics (Chapter 2), most of the work addressing differentia extraction has relied upon manually constructed extraction rules (Vanderwende, 1995; Barrière, 1997; Rus, 2001). Here the emphasis is switched from transformation patterns for extracting relations into statistical classification for relation refinement, given tagged corpora with examples. This allows for better coverage at the expense of precision. Note that, in Extended WordNet (Rus, 2001) relation refinement is not yet addressed: prepositions are converted directly into predicates in the underlying logical form representation (e.g., *by(e, x)*). In addition, more so than the other work, it is tied into the specifics of underlying parser, as a transformation rule is developed for each grammar rule used in parsing.

By isolating the refinement from the syntactic extraction step, this approach can be more readily adapted to related tasks, such as in extracting semantic information from encyclopedia texts. Moreover, this approach is more readily adaptable 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. For instance, there are currently three projects on the creation of FrameNet-style lexicons for languages other than English (German, Spanish, and Japanese).¹

Barnbrook's (2002) definition analysis is also tied in the specifics of a particular dictionary, namely 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

¹See <http://www.icsi.berkeley.edu/framenet/FNabroad.html>.

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. His analysis is more in the vein of language exploration, in particular for the genre of definitions for language learners. 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.

As with the other work in differentia extraction, this thesis addresses the acquisition of conceptual distinctions. In principle, it can handle any level of granularity; however, addressing distinctions at the level of near-synonyms (Edmonds and Hirst, 2002) might require customized analysis for each cluster of nearly synonymous words to be refined. Inkpen and Hirst (2001) discuss how this can be automated 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.

There have been a few papers criticizing work on extracting information from machine-readable dictionaries. Such criticisms do raise some valid points that must be taken in consideration. Amsler (1995) points out 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 also notes that the information might not be suitable for a broad range of uses. Ide and Veronis (1993) provide more details on uniformity issues, in particular with respect to the genus hierarchies. They also point out how differentiating relations are included haphazardly. For instance, in the definitions of 'abricot' (apricot) and 'pêche' (peach) in three different French dictionaries, only peach is indicated as having a hard pit. Nevertheless, work in extraction information from MRD's can be beneficial to the extent that the techniques apply to extracting information 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 little change to the underlying extraction processes.

6.1.2 Relation Refinement

The work here addresses relation refinement specifically with respect to those indicated by prepositional phrases. As discussed earlier, this can be viewed as preposition word-sense disambiguation. Until recently, there has not been much work specifically on preposition classification, especially with respect to general applicability in contrast to 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 all the annotations in 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. However, due to the conditioning of the classification on the predicating word, the range of roles for a particular classification is more limited than in the experiments discussed here.

Blaheta and Charniak (2000) classify semantic role assignments using all the annotations in 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%,² Their tasks are somewhat different, since they address all adjuncts, not just prepositions, hence their lower baseline. In addition, they include the *nominal* and *adverbial* roles, which are syntactic and presumably more predictable than the others in this group. Van den Bosch and Bucholz (2002) also use the Treebank data to address the more general 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 just based on the words that occur in the definitions. As they note, this lacks generalization ability. 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 refinement 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.

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:

²They 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 illustrated 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 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.

6.1.3 Semantic Relatedness

Richardson (1997) discusses how to determine semantic relatedness based on the information extracted from MRD's, using relations extracted using Vanderwende's (1996) techniques. This forms the basis for Microsoft's MindNet system (Richardson et al., 1998) which establishes linkages between arbitrary words and named entities. All the relations extracted from the same definition are first grouped into the same structure and then these structures are inverted to make backward links 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 weight 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 the WordNet have been determined. For example, an upward direction is not allowed following a downward direction segment. In addition, change of direction can only be done once.

6.1.4 Relation Weighting

Richardson (1997) uses a novel procedure is used for weighting the relationships, using notions derived from the weighting of terms in information retrieval by 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. He uses semantic relations in place of terms, so the frequencies are those for the relational triples. 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

The use of cue validity (CV) weights would be roughly equivalent if the set of contrasting concepts corresponded to those in this definition of 'document.' This is because both measures are proportional to the joint frequency with which the feature (term) co-occurs with the class (document). Although the measures differ significantly with respect to the scaling factor applied to the joint frequency, the weights would likely be correlated when normalized. However, the weighting applied here will differ significantly because the reference class against which the joint frequencies are calculated is based only on semantic grounds (e.g., most informative subsumer for the concept being defined) rather than morphological ones (e.g., homonymy of words used to define concepts). In his case, the relationships are among words rather than concepts. He tabulates frequency over the entire relationship but uses smoothing based on the relation/target pair frequencies. In addition, the relation types are weighted to favor those that do not occur too frequently or too infrequently.

6.1.5 Word Sense Disambiguation

Several approaches to word-sense disambiguation (WSD) have relied upon word-overlap in dictionary definitions, starting with (Lesk, 1986). The idea is to select those senses of a target word that have definitions with the most overlap with the definitions for the other words in the sentence. 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. Veronis 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 adding incorporating links from senses to words, and there is no distinguishing of important lexical relations from incidental ones. The probabilistic spreading activation approach discussed in Chapter 5 improves upon

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

word overlap in several respects: for example, it provides differential weighting of the words based on semantic relatedness; and, by using Bayesian Networks, evidence is combined in a sound manner.

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 they are weak techniques useful more as a fallback approach 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 again 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.

Supervised approaches to WSD tend to 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. In addition, although some of the words occurring in dictionary definition are presumably very indicative of a sense, there is no straightforward way to filter out the unrelated definitional 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 independence, only those definitional words related to words that co-occur frequently with the sense are considered.

6.1.6 Class-based Collocations

Scott and Matwin (1998) also use WordNet hypernyms for classification, in particular topic detection. Their approach is different in that they include a numeric density feature for each synset that subsumes words appearing in the document, potentially yielding hundreds of features. The hypernym collocations used here just have a binary feature for each of the relations being classified. In addition to nouns and verbs, adjectives are included as well. The adjective hier-

archy is augmented by treating *is-similar-to* as *has-hypernym*. Adverbs would be included, but there is no hierarchy for them. Adverbs are related to adjectives via *is-derived-from*, so future work might treat these as *has-hypernym*. As with the hypernym collocations, Scott and Matwin 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 results comparable to that obtained via clustering. In contrast, the hypernym collocations used here lead to significant improvement as shown in Section 4.3.

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,

$\langle *, \text{noun}, *, \text{room}_{\text{area}} \rangle \langle \text{'door'}, \text{noun}, \text{door}_{\text{movablebarrier}}, * \rangle$

would match “kitchen door” and “bedroom door”.

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

6.1.7 Bayesian Networks

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 earlier, 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.

6.2 Future Work

The differentia extraction process described in this thesis can be used for other applications besides word-sense disambiguation. This section sketches out a few of these areas. In addition, other areas for future work are discussed.

6.2.1 Application to Text Segmentation

In addition to word-sense disambiguation, the conceptual differentia-extraction work can be applied to text segmentation. The idea is that since related words tend to occur together (Morris and Hirst, 1991), the frequency of related words can serve as an indication of text cohesion and thus can be used to estimate segment boundaries. Hearst's (1994) text segmentation program (TextTiling) serves as a good example for 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 earlier work, Hearst (1993) incorporated thesaural relations into her algorithm. She used Yarowsky's (1992) classifier to assign the most plausible category to each word (using WordNet instead of Roget's). She tabulates frequency using the category label rather than the word itself. She found that this seemed to help with the similarity assessments. However, later work (Hearst, 1994) noted that these relations were no longer incorporated because they led to degraded performance with a revised algorithm.

A direct extension of this approach to one using conceptual differentia would be to assign equivalence classes to words based on occurrence in differentia-based relations. Strictly speaking, these would be considered as semantic-relatedness classes, given that the classes can overlap. Again, the ambiguity of the dictionary differentia pose a problem, so methods will be needed to handle conflicting equivalence class assignments and to avoid having the equivalence classes become too general.

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

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

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.

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

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. In the proposed extension, the frequency counts of the words will be augmented with those for the equivalence classes. A sketch of the revised text segmentation algorithm follows:

Let graph $G = \langle \phi, \phi \rangle$

For each word in text (in order)

 Extract differentia-based relations from dictionary definition

 Optionally, prune the relations based on prior context

 Add relations to graph

Compute connected components in G

Augment each word in text by label of its corresponding component in G

Run TextTiling algorithm over the transformed text

Note that the tokens for the relatedness-based equivalence classes augment the word tokens in the text rather than replacing them. If particular words occur frequently in the text, then it is desirable to retain them as dimensions in the document space. However, for related words that occur infrequently individually, it is useful to incorporate a new dimension for the given class, which would thus have higher weight than the dimensions for the given words. That is, some of the more high-frequency terms will become separate dimensions in the vector space, as well as being incorporated into a dimension with other related terms. To avoid handling this without using arbitrary thresholds, each class label is treated as an additional token, not a replacement to the original word. In effect, this augments the vectors computed by the original algorithm by a component consisting of the frequency counts for class labels:

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

6.2.2 Mapping Senses from other Dictionaries into WordNet

Given that different dictionaries emphasize different aspects of word meaning, it will be desirable to combine the information acquired. This undoubtedly will require human intervention since determining which aspects of the different meanings for the same senses will involve subtle decision making. The 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.

Additional processes not discussed here would be required. 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 correspond. Although simple word-overlap schemes could help in this respect (2001), it would be desirable to integrate work on aligning ontologies (O'Hara et al., 1998; Hovy, 1998) to account for WordNet's hierarchy.

6.2.3 Transferring Semantic Roles across Resources

In Section 4.3, class-based collocations are used partly to overcome the difference in the training data (e.g., FrameNet) and the target data (i.e., definition text). The same approach can be used when training on FrameNet and then testing on Treebank or vice versa. This could be used to verify that mappings between the inventories are sufficient for transferring semantic roles across the corpus resources.

Initial experiments were done training over FrameNet and then testing over Treebank, using both word-based and hypernym-based collocations, as well as their combination. To account for the semantic role differences, mappings were defined to establish rough correspondences between the two datasets. In some cases, no mappings can be established. For example, the FrameNet *interlocutors* role has no corresponding role in Treebank, which does not tag subjects with semantic tags. In addition, there are several cases where the mapping is unclear. For instance, *jurisdiction* can either correspond directly to *locative* or alternatively can specialize the *extent* role. Note that these mappings are different from that discussed in Section 4.3.5, which represents the combination of different role inventories. Because a single inventory is the target, gaps are more likely to occur in this case.

Table 6.1 shows preliminary results for these experiments. The overall results are quite low, due to discrepancies in the inventories. For example, the

| Experiment | Accuracy | # Instances: | 27148 |
|------------|----------|--------------|-------|
| Word Only | 17.3 | # Classes: | 7 |
| Hypernym | 14.6 | Entropy: | 1.847 |
| Combined | 16.0 | Baseline: | 0.491 |

Table 6.1: **Results training over FrameNet and testing over Treebank.** The FrameNet annotations are mapped into the Treebank roles during training. For testing, the Treebank data is used as is. See Table 4.10 for the legend.

most frequent role in Treebank, *Temporal*, occurs in 49% of the testing data instances but occurs less than 5% of the mapped instances in the training data.

6.2.4 Inferring other Types of Relations

In Section 4.3.4.1, relational markers were inferred for the relationships in the Factotum semantic network. Again, 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.

For example, the Cyc KB provides a rich source of attribute relations. In particular, properties for members of a category are specified via the *relation-AllInstance* predicate. A few examples follow:

```
(relationAllInstance numberOfEdges Nonagon 9)
(relationAllInstance objectHasColor Slug BeigeColor)
(relationAllInstance tasteOfObject BakingChocolate-Unsweetened
BitterTaste)
(relationAllInstance hardnessOfObject StoneStuff Hard)
(relationAllInstance mainColorOfObject StoutBeer BlackColor)
```

Note that *relationAllInstance* is a shorthand notation 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 instance assertions in the KB, many of which deal with attribute specifications. For the first example, the relationship would be $\langle \text{Nonagon}, \text{numberOfEdges}, 9 \rangle$. Doing a proximity search on “nonagon NEAR 9” produces hits such as 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, common patterns like “9 sided” and “with 9 sides” would likely emerge. Analyzing the results over the entire set of such *numberOfEdges* assertions will likely produce the following patterns:

with $\langle \text{target} \rangle$ sides

has $\langle \text{target} \rangle$ sides

$\langle \text{target} \rangle$ -sided

6.2.5 Analyzing Lexical Gaps

Bilingual dictionaries are an important resource for machine translation. Usually, the entries just consist of target language words that generally mean the same as the source language word. For instance,⁴

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 cases like these, the entries in bilingual dictionaries give brief definitions more akin to monolingual dictionaries, such as in the following:

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

In such cases, 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.

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 would be 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-STRUCT part of the frame structure while preserving the SEM-STRUCT (see Section 2.1.3.3). The system would then present the suggested foreign language lexicon entry to the user to verify and correct if necessary. However, in the case of lexical gaps, the differentia extraction system could be used as a fallback mechanism to infer relations for the SEM-STRUCT.

⁴These examples are taken from the Spanish-English Dictionary provided with the NTC Languages of the World CD-ROM from the National Textbook Company.

CHAPTER 7 CONCLUSION

This thesis has presented an empirical methodology for extracting differentiating relations (i.e., *differentia extraction*) that is 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 chapter summarizes the thesis to review the important points that were discussed. In addition, several observations resulting from the research are highlighted. These encompass the main contributions of the research and include other insights based on the work.

7.1 Summary of Thesis

This thesis improves upon previous work on extracting information from dictionary definitions by the use of data-driven relation refinement. This incorporates 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 four human judges to be acceptable in quality as compared to manually correctly 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. As seen there, differentiating relations are important for categorization as revealed by research in cognitive psychology. Other support for this 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 in free text is not likely to be sufficiently directed at acquiring differentiating relations.

7.1.2 Approaches for Lexical Acquisition

Chapter 2 presented a background on lexical semantics and illustrated the common techniques that are being used for acquiring lexical semantics. Several different representation approaches 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 finer-grain distinctions would require more emphasis on stylistics and other pragmatic distinctions (Edmonds and Hirst, 2002).

Manual acquisition is commonly done when quality of lexicon entries is critical (Onyshkevych and Nirenburg, 1994; 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). Early definition analysis work has concentrated on extracting the main semantic category for the word being defined (*genus* extraction), such as 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 primarily relied 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 like Longman's Dictionary of Contemporary English (LDOCE), a popular dictionary for computational linguistics research (Procter, 1978). The WordNet definitions are not as uniform as those in LDOCE, but they still tend to follow the classical *genus/differentia* format (Landau, 2001). The first extraction step thus 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 as the source and target word are collapsed into a single tuple that uses the function word in place of the grammatical relations. The last step in the extraction proper 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 most specific to a given source term are weighted highest.

7.1.4 Refinement into Conceptual Relations

Chapter 4 presents the crucial refinement process, which transforms the syntactically oriented relationships into conceptual relationships. For the relation source and target terms, this amounts to word-sense disambiguation (WSD). Several approaches are sketched for WSD in dictionary definitions. Since WordNet is 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 refinement concentrates on resolving 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 traditionally 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. Lastly, to account for the different granularity 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 like FrameNet provide overly specialized relations.

7.1.5 Lexicon Augmentation and Word-sense Disambiguation

Chapter 5 discusses two aspects on how the work can be applied, including detailed evaluations for each. The output provided by the analysis can be used directly to augment existing lexicons, in particular WordNet. To 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. To provide a baseline for the accuracy, part of the output was manually corrected prior to the evaluation. Inter-coder reliability analysis was done 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 application 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. Traditional 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 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 *is-a* relations. 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 thesis work to that done previously. The background chapter covered earlier differentia-extraction work, so this chapter concentrated on relation refinement, 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 will be in using differentia-derived equivalence classes to augment existing approaches for text segmentation. The relation refinement methodology will also be extended to handle modification-type relations. For instance, information will be inferred from Cyc using the techniques developed for Factotum.

7.2 Significance of Research

7.2.1 Empirical Acquisition of Distinctions from Dictionaries

Contrary to what some of the criticisms of machine-readable dictionary (MRD) research might imply (Amsler, 1995; Ide and Veronis, 1993), this type of analysis can still be quite fruitful. This thesis provides an empirical methodology for extracting information from MRD's that is directly extendable to handling other reference resources such as encyclopedias. In addition, more so than earlier work, this approach is readily adaptable to extracting information in foreign languages. 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.

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 on 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 4.18). 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 will 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 these enhanced Bayesian Networks leads to significant improvements in a system for word-sense disambiguation.

7.2.4 Class-based Collocations for Word-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 this thesis work. The hypernym collocations were designed initially for preposition disambiguation, but they have been found useful for WSD in general. 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, as mentioned above.

7.3 Speculations regarding Computational Semantics

In concluding this thesis, some speculations are offered on the future direction of natural language processing (NLP), in particular, with respect to computational semantics. In many respects, the field is less ambitious than it was thirty years ago, when Schank and Wilks were developing systems for natural language understanding (Schank and Abelson, 1977; 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. 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 corpus-based or 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 corpus-based approach 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 dictionary definitions are a special case of indefinite descriptions, it is likely that analysis of the latter will be amenable to similar approaches. It 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-based approaches are feasible in general (Vieira and Poesio, 2000; McShane and Nirenburg, 2002). Therefore, although the techniques described here have only been applied to dictionary definitions, the research actually is a step towards broad-coverage deep understanding.

APPENDICES

APPENDIX A 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 one can afford a particular purchase, you only need to consider whether you have enough money at your disposal, not the various ways with 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. The former generally taste better and are much stronger; however, they usually cost twice as much as regular coffee. Figure A.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.

$$P(BE, RT, LM, ES, JP) = \\ P(JP) \times P(ES|JP) \times P(LM|ES, JP) \times P(RT|ES, JP, LM) \times \\ P(BE|ES, JP, LM, RT)$$

However, by accounting for the independent assumptions represented in the graph, this formula can be simplified as follows:

$$P(BE, RT, LM, ES, JP) = \\ P(JP) \times P(ES) \times P(LM|ES, JP) \times P(RT) \times P(BE|LM, RT)$$

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

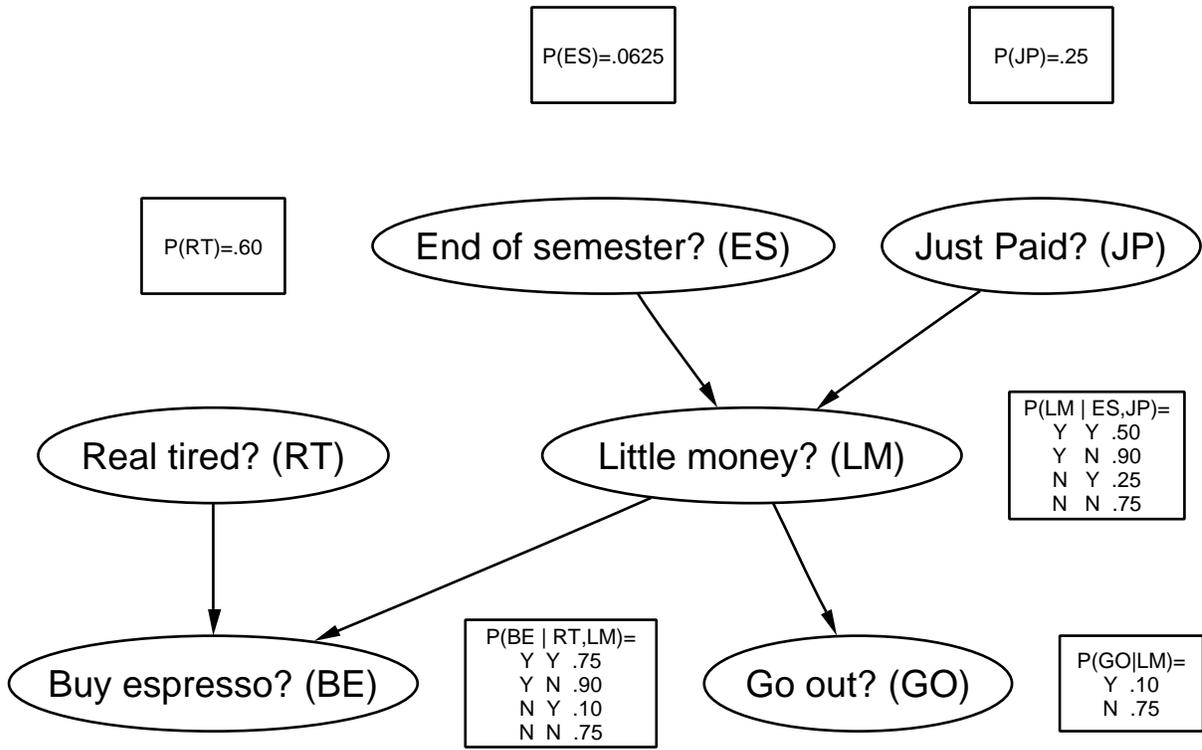


Figure A.1: **Simple Bayesian network for choice of espresso versus regular coffee.**

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*: (*False, False*), (*False, True*), (*True, False*), (*True, True*):

$$P(LM) = 0.75P(\overline{ES})P(\overline{JP}) + 0.25P(\overline{ES})P(JP) + 0.90P(ES)P(\overline{JP}) + 0.50P(ES)P(JP)$$

$$P(LM) = 0.75(1-0.0625)(1-0.25) + 0.25(1-0.0625)(0.25) + 0.90(0.0625)(1-0.25) + 0.50(0.0625)(0.25) = 0.636$$

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 is 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 $P(\overline{JP})$ 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 $P(LM)$ in this case follow:

$$P(LM) = 0.75P(\overline{ES})P(\overline{JP}) + 0.25P(\overline{ES})P(JP) + 0.90P(ES)P(\overline{JP}) + 0.50P(ES)P(JP)$$

$$P(LM) = 0.75(1 - 0.0625)(0) + 0.25(1 - 0.0625)(1) + 0.90(0.0625)(0) + 0.50(0.0625)(1) = 0.266$$

There is a special case that 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.

APPENDIX B PRIMER ON MACHINE LEARNING

This appendix presents a primer for the two machine learning techniques used in the thesis. Both are supervised approaches that rely upon training data providing 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 combination 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 $P(c_i|f_1 \dots f_n)$ is expressed in terms of the probability of the features given the class value $P(f_1 \dots f_n|c_i)$. Specifically,

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

Figure B.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 probability approximation techniques than used in step 2 to account for problems with sparse data.

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:

$$\langle \{f_1, \dots, f_n\}, c_i \rangle$$

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:

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

(Manning and Schütze, 1999; Mitchell, 1997)

Figure B.1: **Methodology for simple Bayesian classification (“Naive Bayes”)**.

| | 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 |
| Frenchfries | 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 B.1: **Training data for favorite-foods examples.**

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

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

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

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

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: $\langle \{F_1, \dots, F_n\}, C_i \rangle$

See Table B.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 $P(C = c_j | F_1 = f_1, \dots, F_n = f_n)$

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

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

Thus enchiladas would be classified as acceptable. The probability of acceptance (i.e., $P(\text{yes} | \dots)$) 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 type of 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 B.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.

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

```
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 B.2: ***Decision tree for favorite-foods example.***

distribution (i.e., yields least uniform distribution). Entropy is a measure of the uniformity of a distribution of values. Higher entropy 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):

$$Entropy = \sum_i -p(x_i) \log_2(p(x_i))$$

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:

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

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 *Fatis* 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 |

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

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

| 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 would lead 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 B.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 alleviate the problem with overfitting of data that decision trees are prone to. 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.

REFERENCES

- Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Joesf Och, David Purdy, Noah A. Smith, and David Yarowsky. 1999. Statistical machine translation: Final report, JHU Workshop 1999. Technical report, Johns Hopkins University.
- Russell G. 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.
- R. Amsler. 1980. *The Structure of the Merriam-Webster Pocket Dictionary*. Ph.D. thesis, University of Texas at Austin.
- Robert A. Amsler. 1995. Introduction. In Guo (Guo, 1995b), pages 1–13.
- B. Atkins. 1995. The dynamic database. In Guo (Guo, 1995b), pages 131–143.
- J. R. Ayto. 1983. On specifying meaning. In R. R. K. Hartmann, editor, *Lexicography: Principles and Practice*, pages 89–98. Academic Press, Inc., London.
- Ken Barker. 1998. *Semi-Automatic Recognition of Semantic Relationships in English Technical Texts*. Ph.D. thesis, Department of Computer Science, University of Ottawa.
- Geoff Barnbrook. 2002. *Defining Language: A Local Grammar of Definition Sentences*. John Benjamins Publishing Company, Amsterdam.
- Caroline Barrière. 1997. *From Machine Readable Dictionaries to a Lexical Knowledge Base of Conceptual Graphs*. Ph.D. thesis, Simon Fraser University.
- Roberto Basili, Maria Teresa Pazienza, and Paolo Verlardi. 1996a. A context driven conceptual clustering method. In B. Boguraev and J. Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*. MIT Press, Cambridge, MA.
- Roberto Basili, Maria Teresa Pazienza, and Paolo Verlardi. 1996b. An empirical symbolic approach to natural language processing. *Artificial Intelligence*, 85:59–99.

Henri Béjoint. 1994. *Tradition and Innovation in Modern English Dictionaries*. Clarendon Press, Oxford.

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for Treebank II style Penn Treebank project. Technical report, University of Pennsylvania.

Don Blaheta and Eugene Charniak. 2000. Assigning function tags to parsed text. In *Proc. NAACL-00*.

B. Boguraev and T. Briscoe, editors. 1989. *Computational Lexicography for Natural Language Processing*. Longman, London.

Ted Briscoe, Ann Copestake, and Alex Lascarides. 1995. Blocking. In Saint-Dizier and Viegas (Saint-Dizier and Viegas, 1995), pages 273–302.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roosin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Peter Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–312.

Rebecca Bruce and Louise Guthrie. 1991. Building a noun taxonomy from a machine readable dictionary. Technical Report MCCS-91-207, Computing Research Laboratory, NMSU.

Rebecca Bruce and Janyce Wiebe. 1999. Decomposable modeling in natural language processing. *Computational Linguistics*, 25(2):195–208.

Bertram Bruce. 1975. Case systems for natural language. *Artificial Intelligence*, 6:327–360.

Kathy J. Burns and Anthony B. 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.

Patrick J. Cassidy. 2000. An investigation of the semantic relations in the Roget's Thesaurus: Preliminary results. In *Proc. CICLing '00*.

- Eugene Charniak. 1992. Bayesian networks without tears. *AI Magazine*, 12(4):50–63.
- E. Charniak. 1993. *Statistical Language Learning*. MIT Press, Cambridge, MA.
- Gennaro Chierchia and Sally McConnell-Ginet. 2000. *Meaning and Grammar*. MIT Press, Cambridge, MA, second edition.
- Timothy Chklovski and Rada 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 *Proc. COLING-92*, pages 359–365. Nantes, France.
- D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.
- I. Dagan, A. Itai, and U. Schwall. 1991. Two languages are more informative than one. In *Proc. 26th Annual Meeting of the Association for Computational Linguistics*. pp. ??-??
- K. Dalgren and J. McDowell. 1986. Using commonsense knowledge to disambiguate prepositional phrase modifiers. In *Proc. 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. Philadelphia, PA, USA.
- Bonnie Dorr, Nizar Habash, and David Traum. 1998. A thematic hierarchy for efficient generation from lexical-conceptual structure. Technical Report 022, UMIACS, University of Maryland.
- Bonnie J. Dorr. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–322.
- David R. Dowty. 1979. *Word Meaning and Montague Grammar*. D. Reidel Publishing, Holland.

- Michael G. Dyer. 1983. *In-depth Understanding*. MIT Press, Cambridge, MA.
- P. Edmonds and S. Cotton, editors. 2001. *Proceedings of the SENSEVAL 2 Workshop*. Association for Computational Linguistics.
- Philip Edmonds and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.
- Phil Edmonds and Adam Kilgarriff. 2002. Special issue on evaluating word sense disambiguation systems. *Journal of Natural Language Engineering*, 8(4). Editors.
- Phil Edmonds. 1999. *Semantic Representations of Near-Synonyms for Automatic Lexical Choice*. Ph.D. thesis, University of Toronto.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Charles J. Fillmore, Charles Wooters, and Collin F. Baker. 2001. Building a large lexical databank which provides deep semantics. In *Proceedings of the Pacific Asian Conference on Language, Information and Computation*. Hong Kong.
- C. Fillmore. 1968. The case for case. In Emmon Bach and Robert T. 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*.
- William Frawley. 1992. *Linguistic Semantics*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
- Pascale Fung and Kenneth Ward Church. 1994. K-vec: A new approach for aligning parallel texts. In *Proc. COLING-94*.
- William A. Gale, Kenneth W. Church, and David 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.
- Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

- Brendan S. Gillon. 1999. The lexical semantics of English count and mass nouns. In Viegas (Viegas, 1999), pages 19–37.
- Cheng-Ming Guo. 1995a. Constructing a MTD from LDOCE. In *Machine Tractable Dictionaries: Design and Construction* (Guo, 1995b).
- Cheng-Ming Guo, editor. 1995b. *Machine Tractable Dictionaries: Design and Construction*. Ablex Publishing Corporation, Norwood, NJ.
- M.A.K. 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 *Proc. of the SIGLEX Workshop*.
- Marti 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 *Proc. 32nd Annual Meeting of the Association for Computational Linguistics*. June 1994.
- 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).
- Irene Heim and Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Blackwell Publishers, Malden, MA.
- Dirk 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.
- Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum (Fellbaum, 1998).
- Graeme Hirst. 1986. Why dictionaries should list case structures. In *Proc. Conference on Advances in Lexicography*. University of Waterloo, November.

- Graeme Hirst. 1988. Resolving lexical ambiguity computationally with spreading activation and polaroid words. In Steven L. Small, Garrison W. Cottrell, and Michael K. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives From Psycholinguistics, Neuropsychology, and Artificial Intelligence*, pages 73–107. Morgan Kaufmann, Los Altos, CA.
- Graeme Hirst. 1995. Near-synonymy and the structure of lexical knowledge. In Klavans (Klavans, 1995), pages 51–56.
- Eduard Hovy. 1998. Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In *Proc. LREC-98*.
- N. Ide and J. Veronis. 1993. Extracting knowledge bases from machine readable dictionaries: Have we wasted our time? In *Proc. KB&KS*.
- 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.
- Diana Inkpen and Graeme Hirst. 2001. Building a lexical knowledge-base of near-synonym differences. In *Proceedings of the Workshop on WordNet and Other Lexical Resources*. NAACL-01.
- R. Jackendoff. 1983. *Semantics and Cognition*. MIT Press, Cambridge, MA.
- R. Jackendoff. 1990. *Semantic Structures*. MIT Press, Cambridge, MA.
- Nathalie Japkowicz and Janyce Wiebe. 1991. Translating spatial prepositions using conceptual information. In *Proc. 29th Annual Meeting of the Assoc. 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.
- Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, New Jersey.
- Mehmet Kayaalp, Ted Pedersen, and Rebecca Bruce. 1997. A statistical decision making method: A case study on prepositional phrase attachment. In *Proc. Computational Natural Language Learning (CoNLL-97)*.
- Adam Kilgarriff and Martha Palmer, editors. 2000a. *Computers and the Humanities: Special Issue on SENSEVAL*, volume 34(1-2). Kluwer Academic Publishers, Dordrecht, the Netherlands.

- Adam Kilgarriff and Martha Palmer. 2000b. Introduction to the special issue on SENSEVAL. In *Computers and the Humanities* (Kilgarriff and Palmer, 2000a), pages 15–48.
- Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for English SENSEVAL. In *Computers and the Humanities* (Kilgarriff and Palmer, 2000a), pages 15–48.
- Adam 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 *Proc. First International Conference on Language Resources and Evaluation*, pages 581–588. Granada, Spain.
- J. Klavans, editor. 1995. *Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*, AAAI Spring Symposium Series, Stanford, March 27-29.
- 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.
- Helen Langone, Benjamin R. Haskell, and George A. Miller. 2004. Annotating wordnet. In *Proc. Frontiers in Corpus Annotation Workshop*. HLT-NAACL 2004.
- S. L. Lauritzen and D. J. 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.
- Geoffrey Leech. 1974. *Semantics*. Middlesex, Penguin Books.
- Fritz 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 *Proc. MUC-4*, pp. 282- 288.

- Douglas B. Lenat and R. V. Guha. 1990. *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project*. Addison-Wesley, Reading, Massachusetts.
- D. B. 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 *Proc. ACM SIGDOC Conference*, pages 24–26. Toronto, Ontario.
- Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.
- Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. COLING-ACL 98*. Montreal.
- Dennis V. Lindley, Glenn Shafer, David J. Spiegelhalter, et al. 1987. Special issue on probability in expert systems. *Statistical Science*, 1(2).
- Kenneth C. Litkowski. 1997. Automatic creation of lexical knowledge bases: New developments in computational lexicology. Technical Report 97-03, CL Research.
- K. C. Litkowski. 2002. Digraph analysis of dictionary preposition definitions. In *Proceedings of the Association for Computational Linguistics Special Interest Group on the Lexicon*. July 11, Philadelphia, PA.
- Rey-Long Liu and Von-Wun Soo. 1993. An empirical study on thematic knowledge acquisition based on syntactic clues and heuristics. In *Proc ACL-93*.
- K. Mahesh and S. Nirenburg. 1995. A situated ontology for practical NLP. In *Proc. Workshop on Basic Ontological Issues in Knowledge Sharing*. International Joint Conference on Artificial Intelligence (IJCAI-95), Aug. 19-20, 1995. Montreal, Canada.
- Christopher D. Manning and Hinrich 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.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proc. ARPA Human Language Technology Workshop*.

Judith Markowitz, Thomas Ahlswede, and Martha Evens. 1986. Semantically significant patterns in dictionary definitions. In *Proc. 24th Annual Meeting of the ACL*, pages 112–119.

James D. McCawley. 1986. What linguists might contribute to dictionary making if they could get their act together. In Peter C. Bjarkman and Victor Raskin, editors, *The Real-World Linguist: Linguistic Applications in the 1980s*, pages 3–18. Ablex, Norwood, NJ.

Marjorie McShane and Sergei Nirenburg. 2002. Reference and ellipsis in ontological semantics. Technical Report MCCS-02-329, Computing Research Laboratory, NMSU.

D. L. Medin, R. L. Goldstone, and D. Gentner. 1993. Respects for similarity. *Psychological Review*, 100:252–278.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–49.

Igor A. Mel'čuk and Alain Polguere. 1987. A formal lexicon in the meaning-text theory (or how to do lexica with words). *Computational Linguistics*, 13(3-4):261–275.

Rada Mihalcea and Dan Moldovan. 2001. A highly accurate bootstrapping algorithm for word sense disambiguation. *International Journal on Artificial Intelligence Tools*, 10(1-2), pages 5–21.

Rada 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)*, Taiwan, August.

George A. 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. (and collected papers), Unpublished MS, August 1993.

- G. Miller, M. Chodorow, S. Landes, C. Leacock, and R. Thomas. 1994. Using a semantic concordance for sense identification. In *Proc. ARPA Human Language Technology Workshop*. San Francisco: Morgan Kaufmann.
- G. Miller. 1990. Introduction. *International Journal of Lexicography*, 3(4): Special Issue on WordNet.
- George A. Miller. 1996. *The Science of Words*. Scientific American Library, New York, second edition.
- Frederick C. Mish, editor. 1996. *Merriam Webster's Collegiate Dictionary*. Merriam-Webster, Incorporated, Springfield, Massachusetts, 10th edition.
- Tom M. Mitchell. 1997. *Machine Learning*. New York, McGraw-Hill.
- Dan I. Moldovan and Vasile Rus. 2001. Logic form transformation of WordNet and its applicability to question answering. In *Proceedings of the ACL 2001 Conference*. July 2001, Toulouse, France.
- Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48.
- Vivi Nastase and Stan Szpakowicz. 2001. Word sense disambiguation in Roget's thesaurus using WordNet. In *Proceedings of the NAACL WordNet and Other Lexical Resources Workshop*, pages 17–22.
- Vivi Nastase and Stan Szpakowicz. 2003. Augmenting WordNet's structure using LDOCE. In *Proc. CICLing, Mexico City, Mexico, February 2003*.
- Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proc. 31st Annual Meeting of the Association for Computational Linguistics*, pages 40–47.
- Sergei Nirenburg and Victor Raskin. 2004. *Ontological Semantics*. MIT Press, Cambridge, MA.
- Adrian Novischi. 2002. Accurate semantic annotations via pattern matching. In *Proceedings of Florida Artificial Intelligence Research Society (FLAIRS 2002)*, Pensacola, Florida.
- Tom O'Hara, Kavi Mahesh, and Sergei Nirenburg. 1998. Lexical acquisition with WordNet and the Mikrokosmos ontology. In *Proc. Usage of WordNet in Natural Language Processing Systems*. COLING-ACL '98 Workshop, August 16, 1998, University of Montreal.

- Tom O'Hara, Janyce Wiebe, and Rebecca F. 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.
- 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. 1994. The lexicon in the scheme of KBMT things. Technical Report MCCS-94-277, Computing Research Laboratory, NMSU.
- B. Onyshkevych and S. Nirenburg. 1995. A lexicon for knowledge-based MT. *Machine Translation*, 10(2):5–57. Special Issue on Building Lexicons for MT.
- OpenCyc. 2002. OpenCyc release 0.6b. <http://www.opencyc.org>.
- Martha Stone Palmer. 1990. *Semantic Processing for Finite Domains*. Cambridge University Press, Cambridge.
- Darren Pearce. 2001. Synonymy in collocation extraction. In *Proc. Workshop on WordNet and Other Lexical Resources*.
- Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.
- T. Pedersen and R. Bruce. 1998. Knowledge-lean word-sense disambiguation. In *Proc. of the 15th National Conference on Artificial Intelligence (AAAI-98)*, pages 800–805. Madison, Wisconsin.
- Ferdnando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proc. 31st Annual Meeting of the Association for Computational Linguistics*.
- William Phillips and Ellen 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, Philadelphia, July. Association for Computational Linguistics.
- P. Procter, editor. 1978. *Longman Dictionary of Contemporary English*. Longman Group, Harlow, Essex.
- J. Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

- M. R. 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. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California.
- Victor Raskin and Sergei Nirenburg. 1995. Lexical semantics of adjectives: A microtheory of adjectival meaning. Technical Report MCCS-95-288, Computing Research Laboratory, NMSU.
- Deepak Ravichandran and Eduard 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.
- Philip 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 *Proc. Third Workshop on Very Large Corpora*. Cambridge, MA, June 1995.
- Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. 1998. MindNet: acquiring and structuring semantic information from text. In *Proc. COLING-98*.
- S. Richardson. 1997. *Determining Similarity and Inferring Relations in a Lexical Knowledge Base*. Ph.D. thesis, City University of New York.
- Ellen Riloff and Mark Schmelzenbach. 1998. An empirical approach to conceptual case frame acquisition. In *Proc. WVLC-98*.
- E. Rosch and C. B. Mervis. 1975. Family resemblance studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.
- Eleanor 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*.
- Vasile Rus. 2002. *Logic Forms for WordNet Glosses*. Ph.D. thesis, Computer Science Department, Southern Methodist University.

- Stuart Russell and Peter 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.
- Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum Associates, Potomac, Maryland.
- Roger C. Schank. 1973. Identification of conceptualizations underlying natural language. In Roger Schank and Kenneth Colby, editors, *Computer Models of Thought and Language*, pages 187–247. W. H. Freeman and Company, San Francisco.
- R. W. Schvaneveldt, D. W. Dearholt, and F. T. Durso. 1988. Graph theoretic foundations of Pathfinder networks. *Computers and Mathematics with Applications*, 15:337–345.
- Sam Scott and Stan Matwin. 1998. Text classification using WordNet hypernyms. In *Proc. Usage of WordNet in Natural Language Processing Systems*, pages 38–44. COLING-ACL '98 Workshop, August 16, 1998, University of Montreal.
- Glenn Shafer. 1976. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey.
- Glenn Shafer. 1987. Probability judgment in artificial intelligence and expert systems. *Statistical Science*, 2:3–44.
- G. Sidorov, I. Bolshakov, P. Cassidy, S. Galicia-Haro, and A. Gelbukh. 1999. 'Non-adult' semantic field: comparative analysis for English, Spanish, and Russian. In *Proc. 3rd Tbilisi Symposium on Language, Logic, and Computation*.
- S. Siegel and N. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. New York, McGraw-Hill, second edition.
- Brian M. Slator and Yorick 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 *Proc. 5th Annual*

Rocky Mountain Conference on Artificial Intelligence (RMCAI-90), pages 257–262.

Daniel Sleator and Davy Temperley. 1993. Parsing English with a link grammar. In *Proc. Third International Workshop on Parsing Technologies*. August 1993.

Steven L. Small and Chuck Rieger. 1982. Parsing and comprehending with word experts (a theory and its realization). In Wendy G. Lehnert and Martin H. Ringle, editors, *Strategies for Natural Language Processing*, Hillsdale, NJ. Lawrence Erlbaum Associates.

Edward E. Smith and Douglas L. Medin. 1981. *Categories and Concepts*. Harvard University Press, Cambridge, MA.

Harold L. Somers. 1987. *Valency and Case in Computational Linguistics*. Edinburgh University Press, Edinburgh, Scotland.

John F. Sowa. 1984. *Conceptual Structures in Mind and Machines*. Addison-Wesley, Reading, MA.

John F. Sowa. 1999. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing, Pacific Grove, CA.

Rohini Srihari, Cheng Niu, and Wei Li. 2001. A hybrid approach for named entity and sub-type tagging. In *Proc. 6th Applied Natural Language Processing Conference*.

Michael 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)*. Arlington, Virginia.

John R. Taylor. 1993. Prepositions: patterns of polysemization and strategies of disambiguation. In Zelinsky-Wibbelt (Zelinsky-Wibbelt, 1993).

Indalecio Arturo Trujillo. 1995. *Lexicalist Machine Translation of Spatial Prepositions*. Ph.D. thesis, University of Cambridge.

Amos Tversky. 1977. Features of similarity. *Psychological Review*, 84(4):327–352.

Henk van Riemsdijk and Edwin Williams. 1986. *Introduction to the Theory of Grammar*. MIT Press, Cambridge, MA.

Lucy Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. Technical Report MSR-TR-94-21, Microsoft Research. Also in *Proc. COLING-94*, August 5-9, 1994, Kyoto, Japan, pp. 782-88.

Lucy Vanderwende. 1995. Ambiguity in the acquisition of lexical information. In Klavans (Klavans, 1995), pages 174–179.

Lucy Vanderwende. 1996. *Understanding Noun Compounds Using Semantic Information Extracted from On-Line Dictionaries*. Ph.D. thesis, Georgetown University.

Jorn Veenstra, Antal van den Bosch, Sabine Buchholz, Walter Daelemans, and Jakub Zavrel. 2000. Memory-based word sense disambiguation. *Computers and the Humanities, special issue on SENSEVAL, Word Sense Disambiguation*, 4(1-2):171–177. Adam Kilgarriff and Martha Palmer, editors.

Jean Veronis and Nancy M. Ide. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proc. COLING-90*. Helsinki, August 1990.

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 *Proc. 31st Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, CA.

Evelyn 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-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. P. Vossen, N. Calzolari, G. Adriaens, A. Sanfilippo, Y. Wilks, editors, Madrid, July 12th, 1997.

Janyce Wiebe, Rebecca Bruce, and Lei Duan. 1997. Probabilistic event categorization. In *Proc. Recent Advances in Natural Language Processing (RANLP-97)*. Tsigov Chark, Bulgaria.

Janyce Wiebe, Kenneth McKeever, and Rebecca Bruce. 1998a. Mapping collocational properties into machine learning features. In *Proc. 6th Workshop on Very Large Corpora (WVLC-98)*, pages 225–233.

- Janyce Wiebe, Tom O'Hara, and Rebecca Bruce. 1998b. Constructing Bayesian networks from WordNet for word-sense disambiguation: Representational and processing issues. In *Proc. Usage of WordNet in Natural Language Processing Systems*, pages 23–30. COLING-ACL '98 Workshop, August 16, 1998, University of Montreal.
- 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.
- Yorick Wilks, Brian M. Slator, and Louise Guthrie. 1996. *Electric Words*. MIT Press, Cambridge, MA.
- Yorick Wilks. 1975a. An intelligent analyzer and understander of English. *CACM*, 18(5):264–274.
- Yorick Wilks. 1975b. A preferential pattern-seeking semantics for natural language inference. *Artificial Intelligence*, 6:53–74.
- Yorick Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11(2):197–223.
- Michael Witbrock, David Baxter, Jon Curtis, Dave Schneider, Robert Kahlert, Pierluigi Miraglia, Peter Wagner, Kathy Panton, Gavin Matthews, and Amanda Vizedom. 2003. An interactive dialogue system for knowledge acquisition in Cyc. In *Proceedings of the Workshop on Mixed-Initiative Intelligent Systems*. IJCAI.
- Ian H. Witten and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA.
- D. Yarowsky, S. Cucerzan, R. Florian, C. Schafer, and R. Wicentowski. 2001. The Johns Hopkins SENSEVAL2 system descriptions. In *Proceedings of SENSEVAL2*, pages 163–166.
- David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proc. COLING-92*, pages 454–460. Nantes, Aug 23-28.
- Cornelia Zelinsky-Wibbelt, editor. 1993. *The Semantics of Prepositions: From Mental Processing to Natural Language Processing*. Mouton de Gruyter, Berlin.