Exploiting Semantic Role Resources for Preposition Disambiguation

Tom O'Hara*
University of Maryland, Baltimore County

Janyce Wiebe**
University of Pittsburgh

This article describes how semantic role resources can be exploited for preposition disambiguation. The main resources include the semantic role annotations provided by the Penn Treebank and FrameNet tagged corpora. The resources also include the assertions contained in the Factotum knowledge base, as well as information from Cyc and Conceptual Graphs. A common inventory is derived from these in support of definition analysis, which is the motivation for this work.

The disambiguation concentrates on relations indicated by prepositional phrases, and is framed as word-sense disambiguation for the preposition in question. A new type of feature for word-sense disambiguation is introduced, using WordNet hypernyms as collocations rather than just words. Various experiments over the Penn Treebank and FrameNet data are presented, including prepositions classified separately versus together, and illustrating the effects of filtering. Similar experimentation is done over the Factotum data, including a method for inferring likely preposition usage from corpora, as knowledge bases do not generally indicate how relationships are expressed in English (in contrast to the explicit annotations on this in the Penn Treebank and FrameNet). Other experiments are included with the FrameNet data mapped into the common relation inventory developed for definition analysis, illustrating how preposition disambiguation might be applied in lexical acquisition.

1. Introduction

English prepositions convey important relations in text. When used as verbal adjuncts, they are the principal means of conveying semantic roles for the supporting entities described by the predicate. Preposition disambiguation is a challenging problem. First, prepositions are highly polysemous. A typical collegiate dictionary has dozens of senses for each of the common prepositions. Second, the senses of prepositions tend to be closely related to one another. For instance, there are three duplicate role assignments among the twenty senses for *of* in The Preposition Project (Litkowski and Hargraves 2006), a resource containing semantic annotations for common prepositions.

* Institute for Language and Information Technologies, Baltimore, MD 21250. E-mail: tomohara@umbc.edu.
** Department of Computer Science, Pittsburgh, PA 15260. E-mail: wiebe@cs.pitt.edu.

Submission received: 7 August 2006; accepted for publication: 21 February 2007.

Consider the disambiguation of the usages of *on* in the following sentences:

(1) The cut should be blocked *on* procedural grounds.

(2) The industry already operates *on* very thin margins.

The choice between the purpose and manner meanings for *on* in these sentences is difficult. The *purpose* meaning seems preferred for sentence 1, as *grounds* is a type of justification. For sentence 2, the choice is even less clear, though the *manner* meaning seems preferred.

This article presents a new method for disambiguating prepositions using information learned from annotated corpora as well as knowledge stored in declarative lexical resources. The approach allows for better coverage and finer distinctions than in previous work in preposition disambiguation. For instance, a traditional approach would involve manually developing rules for *on* that specify the semantic type of objects associated with the different senses (e.g., time for *temporal*). Instead, we infer this based on lexical associations learned from annotated corpora.

The motivation for preposition disambiguation is to support a system for lexical acquisition (O'Hara 2005). The focus of the system is to acquire distinguishing information for the concepts serving to define words. Large-scale semantic lexicons mainly emphasize the taxonomic relations among the underlying concepts (e.g., *is-a* and *part-of*), and often lack sufficient differentiation among similar concepts (e.g., via attributes or functional relations such as *is-used-for*). For example, in WordNet (Miller et al. 1990), the standard lexical resource for natural language processing, the only relations for *beagle* and *Afghan* are that they are both a type of *hound*. Although the size difference can be inferred from the definitions, it is not represented in the WordNet semantic network.

In WordNet, words are grouped into synonym sets called **synsets**, which represent the underlying concepts and serve as nodes in a semantic network. Synsets are ordered into a hierarchy using the **hypernym** relation (i.e., *is-a*). There are several other semantic relations, such as *part-whole*, *is-similar-to*, and *domain-of*. Nonetheless, in version 2.1 of WordNet, about 30% of the synsets for noun entries are not explicitly distinguished from sibling synsets via semantic relations.

To address such coverage problems in lexicons, we have developed an empirical approach to lexical acquisition, building upon earlier knowledge-based approaches in dictionary definition analysis (Wilks, Slator, and Guthrie 1996). This involves a two-step process: Definitions are first analyzed with a broad-coverage parser, and then the resulting syntactic relationships are disambiguated using statistical classification. A crucial part of this process is the disambiguation of prepositions, exploiting online resources with semantic role usage information. The main resources are the Penn Treebank (PTB; Marcus et al. 1994) and FrameNet (Fillmore, Wooters, and Baker 2001), two popular corpora providing rich annotations on English text, such as the semantic roles associated with prepositional phrases in context. In addition to the semantic role annotations from PTB and FrameNet, traditional knowledge bases (KBs) are utilized to provide training data for the relation classification. In particular, the Factotum KB (Cassidy 2000) is used to provide additional training data for prepositions that are used to convey particular relationships. Information on preposition usage is not explicitly encoded in Factotum, so a new corpus analysis technique is employed to infer the associations.

Details on the lexical acquisition process, including application and evaluation, can be found in O'Hara (2005). This article focuses on the aspects of this method relevant to the processing of prepositions. In particular, here we specifically address preposition

disambiguation using semantic role annotations from PTB, FrameNet, and Factotum. In each case, classification experiments are presented using the respective resources as training data with evaluation via 10-fold cross validation.

This article is organized as follows. Section 2 presents background information on the relation inventories used during classification, including one developed specifically for definition analysis. Section 3 discusses the relation classifiers in depth with results given for four different inventories. Section 4 discusses related work in relation disambiguation, and Section 5 presents our conclusions.

2. Semantic Relation Inventories

The representation of natural language utterances often incorporates the notion of *semantic roles*, which are analogous to the slots in a frame-based representation. In particular, there is an emphasis on the analysis of *thematic roles*, which serve to tie the grammatical constituents of a sentence to the underlying semantic representation. Thematic roles are also called case roles, because in some languages the grammatical constituents are indicated by case inflections (e.g., ablative in Latin). As used here, the term "semantic role" refers to an arbitrary semantic relation, and the term "thematic role" refers to a relation intended to capture the semantics of sentences (e.g., event participation).

Which semantic roles are used varies widely in Natural Language Processing (NLP). Some systems use just a small number of very general roles, such as *beneficiary*. At the other extreme, some systems use quite specific roles tailored to a particular domain, such as *catalyst* in the chemical sense.

2.1 Background on Semantic Roles

Bruce (1975) presents an account of early case systems in NLP. For the most part, those systems had limited case role inventories, along the lines of the cases defined by Fillmore (1968). Palmer (1990) discusses some of the more contentious issues regarding case systems, including adequacy for representation, such as reliance solely upon case information to determine semantics versus the use of additional inference mechanisms. Barker (1998) provides a comprehensive summary of case inventories in NLP, along with criteria for the qualitative evaluation of case systems (generality, completeness, and uniqueness). Linguistic work on thematic roles tends to use 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, there has been a trend towards the use of larger sets of semantic primitives (Wilks, Slator, and Guthrie 1996). The WordNet lexicon (Miller et al. 1990) serves as one example of this. A synset is defined in terms of its relations with any of the other 100,000+ synsets, rather than in terms of a set of features like [±ANIMATE]. There has also been a shift in focus from deep understanding (e.g., story comprehension) facilitated by specially constructed KBs to shallow surface-level analysis (e.g., text extraction) facilitated by corpus analysis. Both trends seem to be behind the increase in case inventories in two relatively recent resources, namely FrameNet (Fillmore, Wooters, and Baker 2001) and OpenCyc (OpenCyc 2002), both of which define well over a hundred case roles. However, provided that the case roles are well structured in an inheritance hierarchy, both paraphrasability and coverage can be addressed by the same inventory.

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, following Fillmore (1968). The Berkeley FrameNet (Fillmore, Wooters, and Baker 2001) project currently provides the most comprehensive set of semantic roles annotations. These are at a much finer granularity than those in PTB, making them quite useful for applications learning semantics from corpora. Relation disambiguation experiments for both of these role inventories are presented subsequently.

2.2.1 Penn Treebank. The original PTB (Marcus, Santorini, and Marcinkiewicz 1993) provided syntactic annotations in the form of parse trees for text from the *Wall Street Journal*. This resource is very popular in computational linguistics, particularly for inducing part-of-speech taggers and parsers. PTB version II (Marcus et al. 1994) added 20 functional tags, including a few thematic roles such as *temporal*, *direction*, and *purpose*. These can be attached to any verb complement but normally occur with clauses, adverbs, and prepositions.

For example, Figure 1 shows a parse tree using the extended annotation format. In addition to the usual syntactic constituents such as *NP* and *VP*, function tags are included. For example, the second NP gives the subject. This also shows that the first prepositional phrase (PP) indicates the time frame, whereas the last PP indicates the

Sentence:
In 1982, Sports & Recreation's managers and certain passive investors purchased the company from Brunswick Corp. of Skokie, Ill.

Parse:

(S (PP-**TMP** In (NP 1982)),	*temporal extent*
(NP-**SBJ**	*grammatical subject*
(NP (NP (NP Sports) & (NP Recreation) 's)	
managers)	
and (NP certain passive investors))	
(VP purchased	
(NP the company)	
(PP-**CLR** from	*closely related*
(NP (NP Brunswick Corp.)	
(PP-**LOC** of	*locative*
(NP (NP Skokie) , (NP Ill)))	
))) .)	

Figure 1
Penn Treebank II parse tree annotation sample. The functional tags are shown in **boldface.**

Table 1
Frequency of Penn Treebank II semantic role annotations. Relative frequencies estimated over the counts for unique assignments given in the PTB documentation (bkt_tags.lst), and descriptions based on Bies et al. (1995). Omits low-frequency *benefactive* role. The syntactic role annotations generally have higher frequencies; for example, the *subject* role occurs 49% of the time (out of about 240,000 total annotations).

Role	Freq.	Description
temporal	.113	indicates when, how often, or how long
locative	.075	place/setting of the event
direction	.026	starting or ending location (trajectory)
manner	.021	indicates manner, including instrument
purpose	.017	purpose or reason
extent	.010	spatial extent

location. The second PP is tagged as *closely-related*, which is one of the miscellaneous PTB function tags that are more syntactic in nature: "[CLR] occupy some middle ground between arguments and adjunct" (Bies et al. 1995). Frequency information for the semantic role annotations is shown in Table 1.

2.2.2 FrameNet. FrameNet (Fillmore, Wooters, and Baker 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** (FE). For example, in the *communication* frame, there are frame elements for *communicator*, *message*, *medium*, and so forth. FrameNet annotations occur at the phrase level instead of the grammatical constituent level as in PTB. Figure 2 shows an example.

Table 2 displays the top 25 semantic roles by frequency of annotation. This shows that the semantic roles in FrameNet can be quite specific, as with the roles *cognizer*, *evaluee*, and *addressee*. In all, there are over 780 roles annotated with over 288,000 tagged instances.

Sentence:
Hewlett-Packard Co has rolled out a new range of ISDN connectivity enabling standalone workstations to communicate over public or private ISDN networks.

Annotation:
Hewlett-Packard Co has rolled out a new range of ISDN connectivity enabling
⟨C **FE="Communicator"** PT="NP"⟩standalone workstations⟨/C⟩
to ⟨C TARGET="y"⟩communicate⟨/C⟩
⟨C **FE="Medium"** PT="PP"⟩over public or private ISDN networks⟨/C⟩ .

Figure 2
FrameNet annotation sample. The constituent (C) tags identify the phrases that have been annotated. The frame element (FE) attributes indicate the semantic roles, and the phrase type (PT) attributes indicate the traditional grammatical category for the phrase. For simplicity, this example is formatted in the earlier FrameNet format, but the information is taken from the latest annotations (lu5.xml).

Table 2
Common FrameNet semantic roles. The top 25 of 773 roles are shown, representing nearly half of the total annotations (about 290,000). Descriptions based on FrameNet 1.3 frame documentation.

Role	Freq.	Description
agent	.037	person performing the intentional act
theme	.031	object being acted on, affected, etc.
experiencer	.029	being who has a physical experience, etc.
goal	.028	endpoint of the path
speaker	.028	individual that communicates the message
stimulus	.026	entity that evokes response
manner	.025	manner of performing an action, etc.
degree	.024	degree to which event occurs
self-mover	.023	volitional agent that moves
message	.021	the content that is communicated
path	.020	the trajectory of motion, etc.
cognizer	.018	person who perceives the event
source	.017	the beginning of the path
time	.016	the time at which the situation occurs
evaluee	.016	thing about which a judgment has been made
descriptor	.015	attributes, traits, etc. of the entity
body-part	.014	location on the body of the experiencer
content	.014	situation or state-of-affairs that attention is focused on
topic	.014	subject matter of the communicated message, etc.
item	.012	entity whose scalar property is specified
target	.011	entity which is hit by a projectile
garment	.011	clothing worn
addressee	.011	entity that receives a message from the communicator
protagonist	.011	person to whom a mental property is attributed
communicator	.010	the person who communicates a message

2.3 Other

A recent semantic role resource that is starting to attract interest is the Proposition Bank (PropBank), developed at the University of Pennsylvania (Palmer, Gildea, and Kingsbury 2005). It extends the Penn Treebank with information on verb subcategorization. The focus is on annotating all verb occurrences and all their argument realizations that occur in the *Wall Street Journal*, rather than select corpus examples as in FrameNet. Therefore, the role inventory is heavily verb-centric, for example, with the generic labels *arg0* through *arg4* denoting the main verbal arguments to avoid misinterpretations. Verbal adjuncts are assigned roles based on PTB version II (e.g., argM-LOC and argM-TMP). PropBank has been used as the training data in recent semantic role labeling competitions as part of the Conferences on Computational Natural Language Learning (Carreras and Màrquez 2004, 2005). Thus, it is likely to become as influential as FrameNet in computational semantics.

The Preposition Project similarly adds information to an existing semantic role resource, namely FrameNet. It is being developed by CL Research (Litkowski and Hargraves 2006) and endeavors to provide comprehensive syntactic and semantic information on various usages of prepositions, which often are not represented well in semantic lexicons (e.g., they are not included at all in WordNet). The Preposition Project uses the sense distinctions from the Oxford Dictionary of English and integrates syntactic information about prepositions from comprehensive grammar references.

2.4 Inventories for Knowledge Representation

This section describes three case inventories: one developed for the Cyc KB (Lenat 1995), one used to define Conceptual Graphs (Sowa 1984), and one for the Factotum KB (Cassidy 2000). The first two are based on a 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. Factotum is also based on a knowledge representation paradigm, but in a sense also reflects the empirical aspect of the corpus annotation approach, because the annotations were developed to address the relations implicit in Roget's Thesaurus.

In this article, relation disambiguation experiments are only presented for Factotum, given that the others do not readily provide sufficient training data. However, the other inventories are discussed because each provides relation types incorporated into the inventory used below for the definition analysis (see Section 3.5).

2.4.1 Cyc. The Cyc system (Lenat 1995) is the most ambitious knowledge representation project undertaken to date, in development since 1984. The full Cyc KB is proprietary, which has hindered its adoption in natural language processing. However, to encourage broader usage, portions of the KB have been made freely available to the public. For instance, there is 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. In addition, researchers can obtain access to ResearchCyc, which contains most of the KB except for proprietary information (e.g., internal bookkeeping assertions).

Cyc uses a wide range of role types: very general roles (e.g., *beneficiary*); commonly occurring situational roles (e.g., *victim*); and highly specialized roles (e.g., *catalyst*). Of the 8,756 concepts in OpenCyc, 130 are for event-based roles (i.e., instances of *actor-slot*) with 51 other semantic roles (i.e., other instances of *role*). Table 3 shows the most commonly used event-based roles in the KB.

2.4.2 Conceptual Graphs. The Conceptual Graphs (CG) mechanism was introduced by Sowa (1984) for knowledge representation as part of his Conceptual Structures theory. The original text listed two dozen or so thematic relations, such as *destination* and *initiator*. In all, 37 conceptual relations were defined. This inventory formed the basis for most work in Conceptual Graphs. Recently, Sowa (1999) updated the inventory to allow for better hierarchical structuring and to incorporate the important thematic roles identified by Somers (1987). Table 4 shows a sample of these roles, along with usage estimates based on corpus analysis (O'Hara 2005).

2.4.3 Factotum. The Factotum semantic network (Cassidy 2000) developed by Micra, Inc., makes explicit many of the relations in Roget's Thesaurus.[1] Outside of proprietary resources such as Cyc, Factotum is the most comprehensive KB with respect to *functional relations*, which are taken here to be non-hierarchical relations, excluding attributes. 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.

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

Table 3
Most common event-based roles in OpenCyc. Descriptions based on comments from the OpenCyc knowledge base (version 0.7). Relative frequencies based on counts obtained via Cyc's utility functions.

Role	Freq.	Description
done-by	.178	relates an event to its "doer"
performed-by	.119	doer deliberately does act
object-of-state-change	.081	object undergoes some kind of intrinsic change of state
object-acted-on	.057	object is altered or affected in event
outputs-created	.051	object comes into existence sometime during event
transporter	.044	object facilitating conveyance of transportees
transportees	.044	object being moved
to-location	.041	where the moving object is found when event ends
object-removed	.036	object removed from its previous location
inputs	.036	pre-existing event participant destroyed or incorporated into a new entity
products	.035	object is one of the intended outputs of event
inputs-destroyed	.035	object exists before event and is destroyed during event
from-location	.034	where some moving-object in the move is found at the beginning
primary-object-moving	.033	object is in motion at some point during the event, and this movement is focal
seller	.030	agent sells something in the exchange
object-of-possession-transfer	.030	rights to use object transferred from one agent to another
transferred-thing	.030	object is being moved, transferred, or exchanged in the event transfer
sender-of-info	.030	sender is an agent who is the source of information transferred
inputs-committed	.028	object exists before event and continues to exist afterwards, and as a result of event, object becomes incorporated into something created during event
object-emitted	.026	object is emitted from the emitter during the emission event

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

Figure 3 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. For example, the line with "{{result_of}}" indicates that conversion is the result of transforming, as shown in the semantic relation listing that would be extracted. There are over 400 different relations instantiated in the knowledge base, which has over 93,000 assertions. Some of these are quite specialized (e.g., *has-brandname*). In addition, there are quite a few inverse relations, because most of the relations are not symmetrical. Certain features of the knowledge representation are ignored during the relation extraction used later. For example, relation specifications can have qualifier prefixes, such as an ampersand to indicate that the relationship only sometimes holds.

Table 4
Common semantic roles used in Conceptual Graphs. Inventory and descriptions based on Sowa (1999, pages 502–510). The term *situation* is used in place of Sowa's *nexus* (i.e., "fact of togetherness"), which also covers spatial structures. Freq. gives estimated relative frequencies from O'Hara (2005).

Role	Freq.	Description
agent	.267	entity voluntarily initiating an action
attribute	.155	entity that is a property of some object
characteristic	.080	types of properties of entities
theme	.064	participant involved with but not changed
patient	.061	participant undergoing structural change
location	.053	participant of a spatial situation
possession	.035	entity owned by some animate being
part	.035	object that is a component of some object
origin	.035	source of a spatial or ambient situation
experiencer	.035	animate goal of an experience
result	.032	inanimate goal of an act
instrument	.027	resource used but not changed
recipient	.019	animate goal of an act
destination	.013	goal of a spatial process
point-in-time	.011	participant of a temporal situation
path	.011	resource of a spatial or ambient situation
accompaniment	.011	object participating with another
effector	.008	source involuntarily initiating an action
beneficiary	.008	entity benefiting from event completion
matter	.005	resource that is changed by the event
manner	.005	entity that is a property of some process
source	.003	present at beginning of activity
resource	.003	material necessary for situation
product	.003	present at end of activity
medium	.003	resource for transmitting information
goal	.003	final cause which is purpose or benefit
duration	.003	resource of a temporal process
because	.003	situation causing another situation
amount	.003	a measure of some characteristic

Table 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.[2] The relative frequencies just reflect relationships explicitly labeled in the KB data file. For instance, this does not account for implicit *has-subtype* relationships based on the hierarchical organization of the thesaural groups (e.g., ⟨simple-change, *has-subtype*, conversion⟩). The functional relations are shown in boldface. This excludes the meronym or part-whole relations (e.g., *is-conceptual-part-of*), in line with their classification by Cruse (1986) as hierarchical relations. The reason for concentrating on the functional relations is that these are more akin to the roles tagged in PTB and FrameNet.

The information in Factotum complements WordNet through the inclusion of more functional relations (e.g., non-hierarchical relations such as *uses* and *is-function-of*). For comparison purposes, Table 6 shows the semantic relation usage in WordNet version

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

Original data:

A. ABSTRACT RELATION
...
A6 CHANGE (R140 TO R152)
...
A6.1 SIMPLE CHANGE (R140)
...
A6.1.4 CONVERSION (R144)
#144. Conversion.
N. {{has_subtype(change, R140)}} conversion, transformation.
{{has_case: @R7, initial state, final state}}.
{{has_patient: @R3a, object, entity}}.
{{result_of}} {{has_subtype(process, A7.7)}} converting, transforming.
{{has_subtype}} processing.
transition.

Extracted relationships:

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

Figure 3
Sample data from Factotum. Based on version 0.56 of Factotum.

2.1. As can be seen from the table, the majority of the relations are hierarchical.[3] WordNet 2.1 averages just about 1.1 non-taxonomic properties per concept (including inverses but excluding hierarchical relations such as *has-hypernym* and *is-member-meronym-of*). OpenCyc provides a much higher average at 3.7 properties per concept, although with an emphasis on argument constraints and other usage restrictions. Factotum averages 1.8 properties per concept, thus complementing WordNet in terms of information content.[4]

2.5 Combining the Different Semantic Role Inventories

It is difficult to provide precise comparisons of the five inventories just discussed. This is due both to the different nature of the inventories (e.g., developed for knowledge bases as opposed to being derived from natural language annotations) and due to the way the

3 In WordNet, the *is-similar-to* relation for adjectives can be considered as hierarchical, as it links satellite synsets to heads of adjective clusters (Miller 1998). For example, the satellite synsets for "thirsty" and "rainless" are both linked to the head synset for "dry (vs. wet)."

4 These figures are derived by counting the number of relations excluding the instance and subset ones and then dividing by the number of concepts (i.e., ratio of non-hierarchical relations to concepts). Cyc's comments and lexical assertions are also excluded, as these are implicit in Factotum and WordNet. WordNet's *is-derived-from* relations are omitted as lexical in nature (the figure otherwise would be 1.6).

Table 5
Common Factotum semantic roles. These account for 80% of the instances. **Boldface** relations are used in the experiments (Section 3.4.2).

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

relation listings were extracted (e.g., just including event-based roles from OpenCyc). As can be seen from Tables 2 and 3, FrameNet tends to refine the roles for agents (e.g., *communicator*) compared to OpenCyc, which in contrast has more refinements of the object role (e.g., *object-removed*). The Concept Graphs inventory includes more emphasis on specialization relations than the others, as can be seen from the top entries in Table 4 (e.g., *attribute*).

In the next section, we show how classifiers can be automatically developed for the semantic role inventories just discussed. For the application to dictionary definition analysis, we need to combine the classifiers learned over PTB, FrameNet, and Factotum. 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 would

Table 6
Semantic relation usage in WordNet. Relative frequencies for semantic relations in WordNet (173,570 total instances). This table omits lexical relations, such as the *is-derived-from* relation (71,914 instances). Frequencies based on analysis of database files for WordNet 2.1.

Relation	Freq.	Description
has-hypernym	.558	superset relation
is-similar-to	.130	similar adjective synset
is-member-meronym-of	.071	constituent member
is-part-meronym-of	.051	constituent part
is-pertainym-of	.046	noun that adjective pertains to
is-antonym-of	.046	opposing concept
has-topic-domain	.038	topic domain for the synset
also-see	.019	related entry (for adjectives and verbs)
has-verb-group	.010	verb senses grouped by similarity
has-region-domain	.008	region domain for the synset
has-attribute	.007	related attribute category or value
has-usage-domain	.007	usage domain for the synset
is-substance-meronym-of	.004	constituent substance
entails	.002	action entailed by the verb
causes	.001	action caused by the verb
has-participle	.001	verb participle

have the advantage that new resources could be integrated into the combined relation classifier with minimal effort. However, the resulting role inventory would 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 definition analysis system somewhat unpredictable.

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

The latter approach is employed here. The common inventory incorporates some of the general relation types defined by Gildea and Jurafsky (2002) for their experiments in classifying semantic relations in FrameNet using a reduced relation inventory. They defined 18 relations (including a special-case *null* role for expletives), as shown in Table 7. These roles served as the starting point for the common relation inventory we developed to support definition analysis (O'Hara 2005), with half of the roles used as is and a few others mapped into similar roles. In total, twenty-six relations are defined, including a few roles based on the PTB, Cyc, and Conceptual Graphs inven-

Table 7
Abstract roles defined by Gildea and Jurafsky based on FrameNet. Taken from Gildea and Jurafsky (2002).

agent	cause	degree	experiencer	force	goal
instrument	location	manner	null	path	patient
percept	proposition	result	source	state	topic

Table 8
Inventory of semantic relations for definition analysis. This inventory is inspired by the roles in Table 7 and is primarily based on FrameNet (Fillmore, Wooters, and Baker 2001) and Conceptual Graphs (Sowa 1999); it also includes roles based on the PTB and Cyc inventories.

Relation	Description
accompaniment	entity that participates with another entity
agent	entity voluntarily performing an action
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	non-agentive entity that produces an effect
characteristic	general properties of entities
context	background for situation or predication
direction	either spatial source or goal (same as in PTB)
distance	spatial extent of motion
duration	period of time that the situation applies within
experiencer	entity undergoing some (non-voluntary) experience
goal	location that an affected entity ends up in
instrument	entity or resource facilitating event occurrence
location	reference spatial location for situation
manner	property of the underlying process
means	action taken to affect something
medium	setting in which an affected entity is conveyed
part	component of entity or situation
path	trajectory which is neither a source nor a goal
product	entity present at end of event (same as Cyc *products*)
recipient	recipient of the resource(s)
resource	entity utilized during event (same as Cyc *inputs*)
source	initial position of an affected entity
theme	entity somehow affected by the event
time	reference time for situation

tories. Table 8 shows this role inventory along with a description of each case. In addition to traditional thematic relations, this includes a few specialization relations, which are relevant to definition analysis. 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 type, to subsume other FrameNet roles related to categorization (e.g., *topic*). Note that this inventory is not meant to be definitive and has been developed primarily to address mappings from FrameNet for the experiments discussed in Section 3.5. Thus, it is likely that additional roles will be required when additional sources of semantic relations are incorporated (e.g., Cyc). The mappings were produced manually by reviewing the role descriptions in the FrameNet documentation and checking prepositional usages for each to determine which of the common inventory roles might be most relevant. As some of the roles with the same name have frame-specific meanings, in a few cases this involved conflicting usages (e.g., *body-part* associated with both *area* and *instrument*), which were resolved in favor of the more common usage.[5]

5 See `www.cs.nmsu.edu/~tomohara/cl-prep-article/relation-mapping.html` for the mapping, covering cases occurring at least 50 times in FrameNet.

3. Preposition Disambiguation

This section presents the results of our experiments on the disambiguation of relations indicated by prepositional phrases. Results are given for PTB, FrameNet, and Factotum. The PTB roles are general: For example, for the preposition *for,* there are six distinctions (four, with low-frequency pruning). The PTB role disambiguation experiments thus address a coarse form of sense distinction. In contrast, the FrameNet distinctions are quite specific: there are 192 distinctions associated with *for* (21 with low-frequency pruning); and, there are 17 distinctions in Factotum (15 with low-frequency pruning). Our FrameNet and Factotum role disambiguation experiments thus address fine-grained sense distinctions.

3.1 Overview

A straightforward approach for preposition disambiguation would be to use typical word-sense disambiguation 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 than the application data. For example, the PTB data is from newspaper text (specifically, *Wall Street Journal*), but the lexical acquisition is based on dictionary definitions. We first discuss how class-based collocations address this problem and then present the features used in the experiments.

Before getting into technical details, an informal example will be used to motivate the use of hypernym collocations. Consider the following *purpose* role examples, which are similar to the first example from the introduction.

(3) This contention would justify dismissal of these actions $on_{purpose}$ prudential grounds.

(4) Ramada's stock rose 87.5 cents $on_{purpose}$ the news.

It turns out that *grounds* and *news* are often used as the prepositional object in PTB when the sense for *on* is *purpose* (or reason). Thus, these words would likely be chosen as collocations for this sense. However, for the sake of generalization, it would be better to choose the WordNet hypernym *subject matter*, as that subsumes both words. This would then allow the following sentence to be recognized as indicating purpose even though *censure* was not contained in the training data.

(5) Senator sets hearing $on_{purpose}$ censure of Bush.

3.1.1 Class-Based Collocations via Hypernyms. To overcome data sparseness problems, a class-based approach is used for the collocations, with WordNet synsets as the source of the word classes. (Part-of-speech tags are a popular type of class-based feature used in word sense disambiguation (WSD) to capture syntactic generalizations.) Recall that the WordNet synset hierarchy can be viewed as a taxonomy of concepts. Therefore, in addition to using collocations in the form of other words, we use collocations in the form of semantic concepts.

Word collocation features are derived by making two passes over the training data (e.g., "on" sentences with correct role indicated). The first pass tabulates the

co-occurrence counts for each of the context words (i.e., those in a window around the target word) paired with the classification value for the given training instance (e.g., the preposition sense from the annotation). These counts are used to derive conditional probability estimates of each class value given co-occurrence of the various potential collocates. The words exceeding a certain threshold are collected into a list associated with the class value, making this a "bag of words" approach. In the experiments discussed below, a potential collocate (*coll*) is selected whenever the conditional probability for the class (C) value exceeds the prior probability by a factor greater than 20%:[6]

$$\frac{P(C|coll) - P(C)}{P(C)} \geq .20 \quad (1)$$

That is, for a given potential collocation word (*coll*) to be treated as one of the actual collocation words, the relative percent change of the class conditional probability ($P(C|coll)$) versus the prior probability for the class value ($P(C)$) must be 20% or higher. The second pass over the training data determines the value for the collocational feature of each classification category by checking whether the current context window has any of the associated collocation words. Note that for the test data, only the second pass is made, using the collocation lists derived from the training data.

In generalizing this to a class-based approach, the potential collocational words are replaced with each of their hypernym ancestors from WordNet. The adjective hierarchy is relatively shallow, so it is augmented by treating *is-similar-to* as *has-hypernym*. For example, the synset for "arid" and "waterless" is linked to the synset for "dry (vs. wet)." Adverbs would be included, but there is no hierarchy for them. Because 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 (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 just described is performed over the text consisting of the replacement tokens. Although this introduces noise due to ambiguity, the conditional-probability selection scheme (Wiebe, McKeever, and Bruce 1998) compensates by selecting hypernym synsets that tend to co-occur with specific roles.

6 The 20% threshold is a heuristic that is fixed for all experiments. We tested automatic threshold derivation for Senseval-3 and found that the optimal percentage differed across training sets. As values near 20% were common, it is left fixed rather than adding an additional feature-threshold refinement step.

Note that there is no preference in the system for choosing either specific or general hypernyms. Instead, they are inferred automatically based on the word to be disambiguated (i.e., preposition for these experiments). Hypernyms at the top levels of the hierarchy are less likely to be chosen, as they most likely occur with different senses for the same word (as with *relation#1* previously). However, hypernyms at lower levels tend not to be chosen, as there might not be enough occurrences due to other co-occurring words. For example, *wealth#4* is unlikely to be chosen as a collocation for the second sense of money, as only a few words map into it, unlike *property#2*. The conditional-probability selection scheme (i.e., Equation (1)) handles this automatically without having to encode heuristics about hypernym rank, and so on.

3.1.2 Classification Experiments. A supervised approach for word-sense disambiguation is used following Bruce and Wiebe (1999).

For each experiment, stratified 10-fold cross validation is used: The classifiers are repeatedly trained on 90% of the data and tested on the remainder, with the test sets randomly selected to form a partition. The results described here were obtained using the settings in Figure 4, which are similar to the settings used by O'Hara et al. (2004) in the third Senseval competition. The top systems from recent Senseval competitions (Mihalcea 2002; Grozea 2004) use a variety of lexical features for WSD. Words in the immediate context (*Word*$\pm i$) and their parts of speech (*POS*$\pm i$) are standard features. Word collocations are also common, but there are various ways of organizing collocations into features (Wiebe, McKeever, and Bruce 1998). We use the simple approach of having a single binary feature per sense (e.g., role) that is set true whenever any of the associated collocation words for that sense are encountered (i.e., per-class-binary).

The main difference of our approach from more typical WSD systems (Mihalcea, Chklovski, and Kilgarriff 2004) concerns the hypernym collocations. The collocation context section of Figure 4 shows that word collocations can occur anywhere in the sentence, whereas hypernym collocations must occur within five words of the target

Features:	
Prep:	preposition being classified
POS$\pm$i:	part-of-speech of word at offset *i*
Word$\pm$i:	stem of word at offset *i*
WordColl_r:	context has word collocation for role *r*
$\textit{HypernymColl}_r$:	context has hypernym collocation for role *r*
Collocation context:	
Word:	anywhere in the sentence
Hypernym:	within 5 words of target preposition
Collocation selection:	
Frequency:	$f(word) > 1$
Conditional probability:	$P(C\|coll) \geq .50$
Relative percent change:	$(P(C\|coll) - P(C))/P(C) \geq .20$
Organization:	per-class-binary

Model selection:
C4.5 Decision tree via Weka's J4.8 classifier (Quinlan 1993; Witten and Frank 1999)

Figure 4
Feature settings used in preposition classification experiments. Aspects that differ from a typical WSD system are italicized.

prepositions (i.e., a five-word context window).[7] This reduced window size is used to make the hypernym collocations more related to the prepositional object and the modified term.

The feature settings in Figure 4 are used in three different configurations: word-based collocations alone, hypernym collocations alone, and both collocations together. Combining the two types generally produces the best results, because this balances the specific clues provided by the word collocations with the generalized clues provided by the hypernym collocations.

Unlike the general case for WSD, the sense inventory is the same for all the words being disambiguated; 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. However, if there were sufficient annotations for particular preposition, then it would be advantageous to have a dedicated classifier. For example, the prior probabilities for the roles would be based on the usages for the given preposition. Therefore, we perform experiments illustrating the difference when disambiguating prepositions with a single classifier versus the use of separate classifiers.

3.2 Penn Treebank Classification Experiments

The first set of experiments deals with preposition disambiguation using PTB. When deriving training data from PTB via the parse tree annotations, the functional tags associated with prepositional phrases are converted into preposition sense tags. Consider the following excerpt from the sample annotation for PTB shown earlier:

(6)

(S *(PP-**TMP** In (NP 1982))*,	*temporal extent*
(NP-SBJ	*grammatical subject*
(NP (NP (NP Sports) & (NP Recreation) 's)	
managers) ...	

Treating *temporal* as the preposition sense yields the following annotation:

(7) In_{TMP} 1982, Sports & Recreation's managers ...

The relative frequencies of the roles in the PTB annotations for PPs are shown in Table 9. As can be seen, several of the roles do not occur often with PPs (e.g., *extent*). This somewhat skewed distribution makes for an easier classification task than the one for FrameNet.

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 10 shows the global probabilities for the roles assigned to *at*, along with

7 This window size was chosen after estimating that on average the prepositional objects occur within 2.3 ± 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 PTB.

Table 9
Penn Treebank semantic roles for PPs. Omits low-frequency *benefactive* relation. Freq. is the relative frequency of the role occurrence (36,476 total instances). Example usages are taken from the corpus.

Role	Freq.	Example
locative	.472	workers *at* a factory
temporal	.290	expired *at* midnight Tuesday
direction	.149	has grown *at* a sluggish pace
manner	.050	CDs aimed *at* individual investors
purpose	.030	opened *for* trading
extent	.008	declined *by* 14%

conditional probabilities for these roles given that certain high-level WordNet synsets occur in the context. In a context referring to a concrete concept (i.e., entity#1), the difference in the probability distributions for the *locative* and *temporal* roles shows that the *locative* interpretation becomes even more likely. In contrast, in a context referring to an abstract concept (i.e., abstraction#6), the difference in the probability distributions for the same roles shows that the *temporal* interpretation becomes more likely. Therefore, these class-based lexical associations capture commonsense usages of the preposition *at*.

3.2.2 Results. The classification results for these prepositions in the Penn Treebank show that this approach is very effective. Table 11 shows the accuracy when disambiguating the 14 prepositions using a single classifier with 6 roles. Table 11 also shows the per-class statistics, showing that there are difficulties tagging the *manner* role (e.g., lowest F-score). For the single-classifier case, the overall accuracy is 89.3%, using Weka's J4.8 classifier (Witten and Frank 1999), which is an implementation of Quinlan's (1993) C4.5 decision tree learner.

For comparison, Table 12 shows the results for individual classifiers created for the prepositions annotated in PTB. A few prepositions only have small data sets, such as *of* which is used more for specialization relations (e.g., *category*) than thematic ones. This 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

Table 10
Prior and posterior probabilities of roles for "at" in the Penn Treebank. $P(R)$ is the relative frequency. $P(R|S)$ is the probability of the relation given that the synset occurs in the immediate context of *at*. $RPC_{R,S}$ is the relative percentage change: $(P(R|S) - P(R))/P(R)$.

	Synset				
		entity#1		abstraction#6	
Relation	P(R)	P(R|S)	$RPC_{R,S}$	P(R|S)	$RPC_{R,S}$
locative	73.5	75.5	*0.03*	67.0	−0.09
temporal	23.9	22.5	−0.06	30.6	*0.28*
manner	2.0	1.5	−0.25	2.0	0.00
direction	0.6	0.4	−0.33	0.4	−0.33

Table 11
Overall preposition disambiguation results over Penn Treebank roles. A single classifier is used for all the prepositions. # Instances is the number of role annotations. # Classes is the number of distinct roles. Entropy measures non-uniformity of the role distributions. Baseline is estimated by the most-frequent role. The Word Only experiment uses just word collocations, Hypernym Only just uses hypernym collocations, and Both uses both types of collocations. Accuracy is average for percent correct over ten trials in cross validation. STDEV is the standard deviation over the trials.

Experiment	Accuracy	STDEV
Word Collocations Only	88.1	0.88
Hypernym Collocations Only	88.2	0.43
Both Collocations	89.3	0.33

Data Set Characteristics	
# Instances:	27,308
# Classes:	6
Entropy:	1.831
Baseline:	49.2

	Word Only			Hypernym Only			Both		
Class	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F
direction	.953	.969	.960	.952	.967	.959	.956	.965	.961
extent	.817	.839	.826	.854	.819	.834	.817	.846	.829
locative	.879	.967	.921	.889	.953	.920	.908	.932	.920
manner	.797	.607	.687	.790	.599	.680	.826	.558	.661
purpose	.854	.591	.695	.774	.712	.740	.793	.701	.744
temporal	.897	.776	.832	.879	.794	.834	.845	.852	.848

Table 12
Per-preposition disambiguation results over Penn Treebank roles. A separate classifier is used for each preposition, excluding roles with less than 1% relative frequency. Freq gives the preposition frequency, and Roles the number of senses. Entropy measures data set uniformity, and Baseline selects most common role. The Word and Hypernym columns show results when including just word and hypernym collocations respectively, whereas Both includes both types. Each column shows averages for percent correct over ten trials. The Mean row averages the values of the individual experiments.

Prep	Freq.	Roles	Entropy	Baseline	Word	Hypernym	Both
through	331	4	1.668	0.438	59.795	62.861	58.592
by	1290	7	1.575	0.479	87.736	88.231	86.655
as	220	3	1.565	0.405	95.113	96.377	96.165
between	87	4	1.506	0.483	77.421	81.032	70.456
of	30	3	1.325	0.567	63.182	82.424	65.606
out	76	4	1.247	0.711	70.238	76.250	63.988
for	1401	6	1.189	0.657	82.444	85.795	80.158
on	1915	5	1.181	0.679	85.998	88.720	79.428
in	14321	7	1.054	0.686	86.404	92.647	86.523
throughout	59	2	0.998	0.525	61.487	35.949	63.923
at	2825	5	0.981	0.735	84.178	90.265	85.561
across	78	2	0.706	0.808	75.000	78.750	77.857
from	1521	5	0.517	0.917	91.649	91.650	91.650
to	3074	5	0.133	0.985	98.732	98.537	98.829
Mean	1944.8	4.43	1.12	0.648	80.0	82.1	78.9

experiments. When using preposition-specific classifiers, the hypernym collocations surprisingly outperform the other configurations, most likely due to overfitting with word-based clues: 82.1% versus 80.0% for the word-only case.

3.3 FrameNet Classification Experiments

The second set of experiments perform preposition disambiguation using FrameNet. A similar preposition word-sense disambiguation experiment is carried out over the FrameNet semantic role annotations involving prepositional phrases. Consider the sample annotation shown earlier:

(8) Hewlett-Packard Co has rolled out a new range of ISDN connectivity enabling ⟨C FE="Communicator" PT="NP"⟩standalone workstations⟨/C⟩ to ⟨C TARGET="y"⟩communicate⟨/C⟩ ⟨C ***FE="Medium"*** *PT="PP"⟩over public or private ISDN networks⟨/C⟩*.

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

(9) Hewlett-Packard Co has rolled out a new range of ISDN connectivity enabling standalone workstations to communicate over$_{Medium}$ public or private ISDN networks.

Table 13 shows the distribution of common roles assigned to prepositional phrases. The *topic* role is the most frequent case not directly covered in PTB.

3.3.1 Illustration with "at." See Table 14 for the most frequent roles out of the 124 cases that were assigned to *at*, along with the conditional probabilities for these roles given that certain high-level WordNet synsets occur in the context. In a context referring to concrete entities, the role *place* becomes more prominent. However, in an abstract context, the role *time* becomes more prominent. Thus, similar behavior to that noted for PTB in Section 3.2.1 occurs with FrameNet.

3.3.2 Results. Table 15 shows the results of classification when all of the prepositions are classified together. Due to the exorbitant number of roles (641), the overall results are low. However, the combined collocation approach still shows slight improvement (23.3% versus 23.1%). The FrameNet inventory contains many low-frequency relations

Table 13
Most common FrameNet semantic roles for PPs. Relative frequencies for roles assigned to prepositional phrases in version 1.3 (66,038 instances), omitting cases below 0.01.

Role	Freq.	Role	Freq.	Role	Freq.
goal	.092	theme	.022	whole	.015
path	.071	manner	.021	individuals	.013
source	.043	area	.018	location	.012
topic	.040	reason	.018	ground	.012
time	.037	addressee	.017	means	.011
place	.033	stimulus	.017	content	.011

Table 14
Prior and posterior probabilities of roles for "at" in FrameNet. Only the top 5 of 641 applicable roles are shown. P(R) is the relative frequency for *relation*. P(R|S) is the probability of the relation given that the synset occurs in the immediate context of *at*. $RPC_{R,S}$ is the relative percentage change: (P(R|S) − P(R))/P(R).

	Synset				
		entity#1		abstraction#6	
Relation	P(R)	P(R|S)	$RPC_{R,S}$	P(R|S)	$RPC_{R,S}$
place	15.6	19.0	*21.8*	16.8	*7.7*
time	12.0	11.5	−4.2	15.1	*25.8*
stimulus	6.6	5.0	−24.2	6.6	0.0
addressee	6.1	4.4	−27.9	3.3	−45.9
goal	5.5	6.3	14.5	6.0	9.1

Table 15
Preposition disambiguation with all FrameNet roles. All 641 roles are considered. *Entropy* measures data set uniformity, and *Baseline* selects most common role.

Experiment	Accuracy	STDEV
Word Collocations Only	23.078	0.472
Hypernym Collocations Only	23.206	0.467
Both Collocations	23.317	0.556

Data Set Characteristics	
# Instances:	65,550
# Classes:	641
Entropy:	6.785
Baseline:	9.3

that complicate this type of classification. By filtering out relations that occur in less than 1% of the role occurrences for prepositional phrases, substantial improvement results, as shown in Table 16. Even with filtering, the classification is challenging (e.g., 18 classes with entropy 3.82). Table 16 also shows the per-class statistics, indicating that the *means* and *place* roles are posing difficulties for classification.

Table 17 shows the results when using individual classifiers, ordered by entropy. This illustrates that the role distributions are more complicated than those for PTB, yielding higher entropy values on average. In all, there are over 360 prepositions with annotations, 92 with ten or more instances each. (Several of the low-frequency cases are actually adverbs, such as *anywhere*, but are treated as prepositions during the annotation extraction.) The results show that the word collocations produce slightly better results: 67.8 versus 66.0 for combined collocations. Unlike the case with PTB, the single-classifier performance is below that of the individual classifiers. This is due to the fine-grained nature of the role inventory. When all the roles are considered together, prepositions are sometimes being incorrectly classified using roles that have not been assigned to them in the training data. This occurs when contextual clues are stronger for a commonly used role than for the appropriate one. Given PTB's small role inventory, this problem does not occur in the corresponding experiments.

3.4 Factotum Classification Experiments

The third set of experiments deals with preposition disambiguation using Factotum. Note that Factotum does not indicate the way the relationships are expressed in English.

Table 16
Overall results for preposition disambiguation with common FrameNet roles. Excludes roles with less than 1% relative frequency. *Entropy* measures data set uniformity, and *Baseline* selects most common role. Detailed per-class statistics are also included, averaged over the 10 folds.

Experiment	Accuracy	STDEV
Word Collocations Only	73.339	0.865
Hypernym Collocations Only	73.437	0.594
Both Collocations	73.544	0.856

Data Set Characteristics	
# Instances:	32974
# Classes:	18
Entropy:	3.822
Baseline:	18.4

	Word Only			Hypernym Only			Both		
Class	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F
addressee	.785	.332	.443	.818	.263	.386	.903	.298	.447
area	.618	.546	.578	.607	.533	.566	.640	.591	.613
content	.874	.618	.722	.895	.624	.734	.892	.639	.744
goal	.715	.766	.739	.704	.778	.739	.703	.790	.743
ground	.667	.386	.487	.684	.389	.494	.689	.449	.541
individuals	.972	.947	.959	.961	.945	.953	.938	.935	.936
location	.736	.524	.610	.741	.526	.612	.815	.557	.660
manner	.738	.484	.584	.748	.481	.584	.734	.497	.591
means	.487	.449	.464	.562	.361	.435	.524	.386	.441
path	.778	.851	.812	.777	.848	.811	.788	.849	.817
place	.475	.551	.510	.483	.549	.513	.474	.576	.519
reason	.803	.767	.784	.777	.773	.774	.769	.714	.738
source	.864	.980	.918	.865	.981	.919	.860	.978	.915
stimulus	.798	.798	.797	.795	.809	.802	.751	.752	.750
theme	.787	.811	.798	.725	.847	.779	.780	.865	.820
time	.585	.665	.622	.623	.687	.653	.643	.690	.664
topic	.831	.836	.833	.829	.842	.835	.856	.863	.859
whole	.818	.932	.871	.807	.932	.865	.819	.941	.875

Similarly, WordNet does not indicate this, but it does include definition glosses. For example,

(10)
Factotum:
⟨drying, *is-function-of*, drier⟩

WordNet:
dry_{alter} remove the moisture from and make dry
$dryer_{\text{appliance}}$ an appliance that removes moisture

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

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 from the relationship (i.e., ⟨source,

Table 17
Per-preposition disambiguation results over FrameNet roles. A separate classifier is used for each preposition, excluding roles with less than 1% relative frequency. Freq gives the preposition frequency, and Roles the number of senses. Entropy measures data set uniformity, and Baseline selects most common role. The Word and Hypernym columns show results when including just word and hypernym collocations, respectively, whereas Both includes both types. Each column shows averages for percent correct over ten trials. The Mean row averages the values of the individual experiments.

Prep	Freq.	Roles	Entropy	Baseline	Word	Hypernym	Both
with	3758	25	4.201	19.6	59.970	57.809	61.924
of	7339	22	4.188	12.8	85.747	84.663	85.965
between	675	23	4.166	11.4	61.495	56.215	53.311
under	286	26	4.045	25.5	29.567	33.040	33.691
against	557	26	4.028	21.2	53.540	58.885	31.892
for	2678	22	3.988	22.6	58.135	58.839	39.809
by	3348	18	3.929	13.6	62.618	60.854	61.152
on	3579	22	3.877	18.1	61.011	57.671	60.838
at	2685	21	3.790	21.2	61.814	58.501	57.630
in	6071	18	3.717	18.7	54.253	49.953	53.880
as	1123	17	3.346	27.1	53.585	47.186	42.722
to	4741	17	3.225	36.6	71.963	77.751	72.448
behind	254	13	3.222	22.8	47.560	41.045	43.519
over	1157	16	3.190	27.8	47.911	48.548	50.337
after	349	16	2.837	45.8	62.230	65.395	61.944
around	772	15	2.829	45.1	52.463	52.582	49.357
from	3251	14	2.710	51.2	73.268	71.934	75.423
round	389	12	2.633	34.7	46.531	50.733	49.393
into	1923	14	2.208	62.9	79.175	77.366	80.846
during	242	10	2.004	63.6	71.067	75.200	68.233
like	570	9	1.938	62.3	82.554	79.784	85.666
through	1358	10	1.905	66.0	77.800	77.798	79.963
up	745	10	1.880	60.3	76.328	76.328	74.869
off	647	9	1.830	63.8	90.545	86.854	90.423
out	966	8	1.773	60.7	77.383	79.722	78.671
across	894	11	1.763	67.6	80.291	80.095	80.099
towards	673	10	1.754	67.9	65.681	71.171	65.517
down	965	7	1.600	63.2	81.256	81.466	79.141
along	723	9	1.597	72.5	87.281	86.862	86.590
about	1894	8	1.488	72.2	83.214	76.663	83.899
back	405	7	1.462	64.7	88.103	91.149	86.183
past	275	9	1.268	78.9	85.683	86.423	85.573
Mean	1727.9	14.8	2.762	43.8	67.813	67.453	65.966

relation, target⟩). For example, using AltaVista's Boolean search,[8] this can be done via "source NEAR target."

Unfortunately, this technique would require detailed post-processing of the Web search results, possibly including parsing, in order to extract the patterns. As an expedient, common prepositions[9] are included in a series of proximity searches to find

8 AltaVista's Boolean search is available at `www.altavista.com/sites/search/adv`.
9 The common prepositions are determined from the prepositional phrases assigned functional annotations in the Penn Treebank (Marcus et al. 1994).

the preposition occurring most frequently with the given terms. For instance, given the relationship ⟨drying, *is-function-of*, drier⟩, the following searches would be performed.

(11) drying NEAR drier NEAR in
drying NEAR drier NEAR to
...
drying NEAR drier NEAR "around"

To account for prepositions that occur frequently (e.g., *of*), pointwise mutual information (MI) statistics (Manning and Schütze 1999, pages 66–68) are used in place of the raw frequency when rating the potential markers. These are calculated as follows:

$$MI_{prep} = log_2 \frac{P(X,Y)}{P(X) \times P(Y)} \approx log_2 \frac{f(\text{source NEAR target NEAR prep})}{f(\text{source NEAR target}) \times f(\text{prep})} \quad (2)$$

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

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

Figure 5 gives the feature settings used in the experiments. This is a version of the feature set used in the PTB and FrameNet experiments (see Figure 4), simplified to account for the lack of sentential context. Figure 6 contains sample feature specifications from the experiments discussed in the next section. The top part shows the original relationships from Factotum; the first example indicates that *connaturalize* causes *similarity*. Also included is the most likely relation marker inferred for each instance. This shows that "n/a" is used whenever a preposition for a particular relationship cannot be inferred. This happens in the first example because *connaturalize* is a rare term.

The remaining parts of Figure 6 illustrate the feature values that would be derived for the three different experiment configurations, based on the inclusion of word and/or hypernym collocations. In each case, the classification variable is given by *relation*. For brevity, the feature specification only includes collocation features for the most frequent relations. Sample collocations are also shown for the relations (e.g., *vulgarity* for *is-caused-by*). In the word collocation case, the occurrence of *similarity* is used to determine that the *is-caused-by* feature (WC_1) should be positive (i.e., "1") for the first two instances. Note that there is no corresponding hypernym collocation due to conditional probability filtering. In addition, although *new* is not included as a word collocation, one of its hypernyms, namely *Adj:early#2*, is used to determine that the *has-consequence* feature (HC_3) should be positive in the last instance.

Context:
Source and target terms from relationship (⟨source, *relation*, target⟩)
Features:

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

Collocation selection:

Frequency:	$f(word) > 1$
Relative percent change:	$(P(C\|coll) - P(C))/P(C) \geq .20$
Organization:	per-class-binary grouping

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

Figure 5
Features used in Factotum role classification experiments. Simplified version of Figure 4: Context only consists of the source and target terms.

3.4.3 Results. To make the task more similar to the PTB and FrameNet cases covered previously, only the functional relations in Factotum are used. These are determined by removing the hierarchical relations (e.g., *has-subtype* and *has-part*) along with the attribute relations (e.g., *is-property-of*). In addition, in cases where there are inverse functions (e.g., *causes* and *is-caused-by*), the most frequently occurring relation of each inverse pair is used. This is done because the relation marker inference approach does not account for argument order. The boldface relations in the listing shown earlier in Table 5 are those used in the experiment. Only single-word source and target terms are considered to simplify the WordNet hypernym lookup (i.e., no phrasals). The resulting data set has 5,959 training instances. The data set also includes the inferred relation markers (e.g., one preposition per training instance), thus introducing some noise.

Figure 6 includes a few examples from this data set. This shows that the original relationship ⟨similarity, *is-caused-by*, rhyme⟩ from Factotum is augmented with the *by* marker prior to classification. Again, these markers are inferred via Web searches involving the terms from the original relationship.

Table 18 shows the results of the classification. The combined use of both collocation types achieves the best overall accuracy at 71.2%, which is good considering that the baseline of always choosing the most common relation (*is-caused-by*) is 24.2%. This combination generalizes well by using hypernym collocations, while retaining specificity via word collocations. The classification task is difficult, as suggested by the number of classes, entropy, and baseline values all being comparable to the filtered FrameNet experiment (see Table 16).

3.5 Common Relation Inventory Classification Experiments

The last set of experiments investigate preposition disambiguation using FrameNet mapped into a reduced semantic role inventory. For the application to lexical acquisition, the semantic role annotations are converted into the common relation inventory discussed in Section 2.5. To apply the common inventory to the FrameNet data, annotations using the 641 FrameNet relations (see Table 2) need to be mapped into those

Relationships from Factotum with inferred markers:

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

Word collocations only:

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

Sample collocations:

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

Hypernym collocations only:

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

Sample collocations:

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

Both collocations:

Relation	POS_s	POS_t	Prep	WC_1	...	WC_7	HC_1	HC_2	HC_3	...
is-caused-by	NN	VB	n/a	1	...	0	0	0	0	...
is-caused-by	NN	NN	by	1	...	0	0	0	0	...
has-consequence	NN	JJ	because	0	...	0	0	0	0	...
has-consequence	JJ	VBN	with	0	...	0	0	0	1	...

Legend:

POS_s & POS_t are the parts of speech for the source and target terms; and WC_r & HC_r are the word and hypernym collocations as follows:

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

Figure 6

Sample feature specifications for Factotum experiments. Each relationship from Factotum is augmented with one relational marker inferred via Web searches, as shown at top of figure. Three distinct sets of feature vectors are shown based on the type of collocation included, omitting features for low-frequency relations.

Table 18
Functional relation classification over Factotum. This uses the relational source and target terms with inferred prepositions. The accuracy figures are averages based on 10-fold cross validation. The gain in accuracy for the combined experiment versus the word experiment is statistically significant at $p < .01$ (via a paired t-test).

Experiment	Accuracy	STDEV
Word Collocations Only	68.4	1.28
Hypernym Collocations Only	53.9	1.66
Both Collocations	*71.2*	1.78

Data Set Characteristics	
# Instances:	5,959
# Classes:	21
Entropy:	3.504
Baseline:	24.2

using the 26 common relations shown in Table 8. Results for the classification of the FrameNet data mapped into the common inventory are shown in Table 19. As can be seen, the performance is well above that of the full classification over FrameNet without filtering (see Table 15). Although the low-frequency role filtering yields the highest performance (see Table 16), this comes at the expense of having half of the training instances discarded. Corpus annotations are a costly resource, so such waste is undesirable. Table 19 also shows the per-class statistics, indicating that the *means*, *direction*, and *part* roles are handled poorly by the classifier. The latter two are due to the relatively small training examples for the roles in question, which can be addressed partly by refining the mapping from FrameNet. However, problems classifying the *means* role occur with all classifiers discussed in this article, suggesting that that role is too subtle to be classified with the feature set currently used.

The results in Table 19 also illustrate 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 definition analysis (e.g., *cognizer* and *addressee*). The common inventory therefore strikes a balance between the overly general roles in PTB, 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 definitions. Note that, by using the annotations from PTB and FrameNet, the end result is a general-purpose classifier, not one tied into dictionary text. Thus, it is useful for other tasks besides definition analysis.

This classifier was used to disambiguate prepositions in the lexical acquisition system we developed at NMSU (O'Hara 2005). Evaluation of the resulting distinctions was performed by having the output of the system rated by human judges. Manually corrected results were also evaluated by the same judges. The overall ratings are not high in both cases, suggesting that some of the distinctions being made are subtle. For instance, for "counterintelligence achieved by deleting any information of value" from the definition of *censoring*, *means* is the preferred role for *by*, but *manner* is acceptable. Likewise, *characteristic* is the preferred role for *of*, but *category* is interpretable. Thus, the judges differed considerably on these cases. However, as the ratings for the uncorrected output were close to those for the corrected output, the approach is promising to use for lexical acquisition. If desired, the per-role accuracy results shown in Table 19 could be incorporated as confidence values assigned to particular relationships extracted from definitions (e.g., 81% for those with *source* but only 21% when *means* used).

4. Related Work

The main contribution of this article concerns the classification methodology (rather than the inventories for semantic roles), so we will only review other work related to this aspect. First, we discuss similar work involving hypernyms. Then, we address preposition classification proper.

Scott and Matwin (1998) use WordNet hypernyms for text classification. They include a numeric density feature for any synset that subsumes words appearing in the document, potentially yielding hundreds of features. In contrast, the hypernym collocations discussed in Section 3.1.1 involve a binary feature for each of the relations being classified, using indicative synsets based on the conditional probability test. This test alleviates the need for their maximum height parameter to avoid overly general hypernyms. Their approach, as well as ours, considers all senses of a word, distributing the alternative readings throughout the set of features. In comparison, Gildea

Table 19
Results for preposition disambiguation with common roles. The FrameNet annotations are mapped into the common inventory from Table 8. *Entropy* measures data set uniformity, and *Baseline* selects most common role. Detailed per-class statistics are also included, averaged over the 10 folds.

Experiment	Accuracy	STDEV
Word Collocations Only	62.9	0.345
Hypernym Collocations Only	62.6	0.487
Both Collocations	63.1	0.639

Data Set Characteristics	
# Instances:	59,615
# Classes:	24
Entropy:	4.191
Baseline:	12.2

	Word Only			Hypernym Only			Both		
Class	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F
accompaniment	.630	.611	.619	.671	.605	.636	.628	.625	.626
agent	.623	.720	.667	.639	.726	.677	.616	.731	.668
area	.546	.475	.508	.541	.490	.514	.545	.501	.522
category	.694	.706	.699	.695	.700	.697	.714	.718	.716
cause	.554	.493	.521	.569	.498	.531	.540	.482	.509
characteristic	.595	.468	.523	.607	.474	.530	.584	.490	.532
context	.569	.404	.472	.577	.388	.463	.568	.423	.485
direction	.695	.171	.272	.701	.189	.294	.605	.169	.260
duration	.601	.465	.522	.589	.445	.503	.596	.429	.497
experiencer	.623	.354	.449	.606	.342	.435	.640	.378	.474
goal	.664	.683	.673	.662	.674	.668	.657	.680	.668
instrument	.406	.339	.367	.393	.337	.360	.405	.370	.385
location	.433	.557	.487	.427	.557	.483	.417	.553	.475
manner	.493	.489	.490	.483	.478	.479	.490	.481	.485
means	.235	.183	.205	.250	.183	.210	.254	.184	.212
medium	.519	.306	.382	.559	.328	.412	.529	.330	.403
part	.539	.289	.368	.582	.236	.323	.526	.301	.380
path	.705	.810	.753	.712	.813	.759	.706	.795	.748
product	.837	.750	.785	.868	.739	.788	.769	.783	.770
recipient	.661	.486	.559	.661	.493	.563	.642	.482	.549
resource	.613	.471	.530	.614	.458	.524	.618	.479	.539
source	.703	.936	.802	.697	.936	.799	.707	.937	.806
theme	.545	.660	.596	.511	.661	.576	.567	.637	.600
time	.619	.624	.621	.626	.612	.619	.628	.611	.619

and Jurafsky (2002) instead just select the first sense for their hypernym features for relation classification. They report marginal improvements using the features, whereas configurations with hypernym collocations usually perform best in our preposition disambiguation experiments.

Mohit and Narayanan (2003) use WordNet hypernyms to generalize patterns for information extraction inferred from FrameNet annotations by distributing support from terms co-occurring in annotations for frame elements to the terms for hypernyms. However, they do not incorporate a filtering stage, as with our conditional probability test. 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, as well as their corresponding synsets if known (e.g., sense tagged or unambiguous), and similarly the hypernym synsets if known. To disambiguate a word, the patterns for each of its senses are evaluated in the context, and the sense with the most support is chosen.

The work here addresses relation disambiguation specifically with respect to those indicated by prepositional phrases (i.e., preposition word-sense disambiguation). Until recently, there has been little work on general-purpose preposition disambiguation. Litkowski (2002) and Srihari, Niu, and Li (2001) present approaches using manually derived rules. Both approaches account for only a handful of prepositions; in contrast, for FrameNet we disambiguate 32 prepositions via individual classifiers and over 100 prepositions via the combined classifier. Liu and Soo (1993) present a heuristic approach for relation disambiguation relying upon syntactic clues as well as occurrence of specific prepositions. They assign roles to constituents of a sentence from corpus data provided that sufficient instances are available. Otherwise, a human trainer is used to answer questions needed by the system for the assignment. They report an 86% accuracy rate for the assignment of roles to verbal arguments in about 5,000 processed sentences. Alam (2004) sketches out how the preposition *over* might be disambiguated into one of a dozen roles using features based on the head and complement, such as whether the head is a movement verb or whether the complement refers to a duration. These features form the basis for a manually-constructed decision tree, which is interpreted by hand in an evaluation over sentences from the British National Corpus (BNC), giving a precision of 93.5%. Boonthum, Toida, and Levinstein (2006), building upon the work of Alam, show how WordNet can be used to automate the determination of similar head and complement properties. For example, if both the head and complement refer to people, *with* should be interpreted as accompaniment. These features form the basis for a disambiguation system using manually constructed rules accounting for ten commonly occurring prepositions. They report a precision of 79% with a recall of 76% over an inventory of seven roles in a post hoc evaluation that allows for partial correctness.

There have been a few machine-learning approaches that are more similar to the approach used here. Gildea and Jurafsky (2002) perform relation disambiguation using the FrameNet annotations as training data. They include lexical features for the headword of the phrase and the predicating word for the entire annotated frame (e.g., the verb corresponding to the frame under which the annotations are grouped). They also use several features derived from the output of a parser, such as the constituent type of the phrase (e.g., NP), the grammatical function (e.g., subject), and a path feature listing part-of-speech tags from the target word to the phrase being tagged. They report an accuracy of 78.5% with a baseline of 40.6% over the FrameNet semantic roles. However, by conditioning the classification on the predicating word, the range of roles for a particular classification instance is more limited than in the experiments presented in this article.

Blaheta and Charniak (2000) use the PTB annotations for relation disambiguation. 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 (because prepositions are considered as the head by their parser). As their classifier tags all adjuncts, they include the *nominal* and *adverbial* roles, which are syntactic and more predictable than the roles occurring with prepositional phrases.

There have been recent workshops featuring competitions for semantic role tagging (Carreras and Màrquez 2004, 2005; Litkowski 2004). A common approach is to tag all the semantic roles in a sentence at the same time to account for dependencies, such as via Hidden Markov Models. To take advantage of accurate Support Vector Machine classification, Pradhan et al. (2005) instead use a postprocessing phrase based on trigram models of roles. Their system incorporates a large variety of features, building upon several different preceding approaches, such as including extensions to the path features from Gildea and Jurafsky (2002). Their lexical features include the predicate root word, headwords for the sentence constituents and PPs, as well as their first and last words. Koomen et al. (2005) likewise use a large feature set. They use an optimization phase to maximize satisfaction of the constraints imposed by the PropBank data set, such as the number of arguments for particular predicates (e.g., just two for *stalk, arg0* and *arg1*).

Lastly, Ye and Baldwin (2006) show how filtering can be used to constrain the hypernyms selected to serve as collocations, building upon our earlier work (O'Hara and Wiebe 2003). They report 87.7% accuracy in a setup similar to ours over PTB (i.e., a gain of 2 percentage points). They use a different type of collocation feature than ours: having a binary feature for each potential collocation rather than a single feature per class. That is, they use *Over-Range Binary* rather than *Per-Class Binary* (Wiebe, McKeever, and Bruce 1998). Moreover, they include several hundred of these features, rather than our seven (*benefactive* previously included), which is likely the main source of improvement. Again, the per-class binary organization is a bag of words approach, so it works well only with a limited number of potential collocations. Follow-up work of theirs (Ye and Baldwin 2007) fared well in the recent preposition disambiguation competition, held as part of SemEval-2007 (Litkowski and Hargraves 2007). Thus, an immediate area for future work will be to incorporate such improved feature sets. We will also investigate addressing sentential role constraints as in general semantic role tagging.

5. Conclusion

This article shows how to exploit semantic role resources for preposition disambiguation. Information about two different types of semantic role resources is provided. The emphasis is on corpus-based resources providing annotations of naturally occurring text. The Penn Treebank (Marcus et al. 1994) covers general roles for verbal adjuncts and FrameNet (Fillmore, Wooters, and Baker 2001) includes a wide range of domain-specific roles for all verbal arguments. In addition, semantic role inventories from knowledge bases are investigated. Cyc (Lehmann 1996) provides fine-grained role distinctions, Factotum (Cassidy 2000) includes a variety of functional relations, and work in Conceptual Graphs (Sowa 1999) emphasizes roles for attributes. Relations from both types of resources are considered when developing the inventory of relations used for definition analysis, as shown in Table 8.

The disambiguation concentrates on relations indicated by prepositional phrases, and is framed as word-sense disambiguation for the preposition in question. A new

type of feature for word-sense disambiguation is introduced, using WordNet hypernyms as collocations rather than just words, as is typically done. The full feature set is shown in Figure 4. Various experiments over the PTB and FrameNet data are presented, including prepositions classified separately versus together, and illustrating the effects of filtering. The main results in Tables 11 and 16 show that the combined use of word and hypernym collocations generally achieves the best performance. 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, with results shown in Table 18. In addition, to account for granularity differences in the semantic role inventories, the relations are mapped into a common inventory that was developed based on the inventories discussed in the article. This allows for improved classification in cases where inventories provide overly specialized relations, such as those in FrameNet. Classification results are shown in Table 19.

The recent competitions on semantic relation labeling have highlighted the usefulness of incorporating a variety of clues for general-purpose relation disambiguation (Carreras and Màrquez 2005). Some of the techniques developed here for preposition disambiguation can likely help with relation disambiguation in general. For instance, there are quite a few lexical features, such as in Pradhan et al. (2005), which could be extended to use semantic classes as with our hypernym collocations. In general it seems that, when lexical features are used in supervised machine learning, it is likely that corresponding class-based features based on hypernyms can be beneficial for improved coverage.

Other aspects of this approach are geared specifically to our goal of supporting lexical acquisition from dictionaries, which was the motivation for the emphasis on preposition disambiguation. Isolating the preposition annotations allows the classifiers to be more readily tailored to definition analysis, especially because predicate frames are not assumed as with other FrameNet relation disambiguation. Future work will investigate combining the general relation classifiers with preposition disambiguation classifiers, such as is done in Ye and Baldwin (2006). Future work will also investigate improvements to the application to definition analysis. Currently, FrameNet roles are always mapped to the same common inventory role (e.g., *place* to *location*). However, this should account for the frame of the annotation and perhaps other context information. Lastly, we will also look for more resources to exploit for preposition disambiguation (e.g., ResearchCyc).

Acknowledgments

The experimentation for this article was greatly facilitated though the use of computing resources at New Mexico State University. We are also grateful for the extremely helpful comments provided by the anonymous reviewers.

References

Alam, Yukiko Sasaki. 2004. Decision trees for sense disambiguation of prepositions: Case of over. In *Proceedings of the Computational Lexical Semantics Workshop*, pages 52–59, Boston, MA.

Barker, Ken. 1998. *Semi-Automatic Recognition of Semantic Relationships in English Technical Texts*. Ph.D. thesis, Department of Computer Science, University of Ottawa.

Bies, Ann, 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 MS-CIS-95-06, University of Pennsylvania.

Blaheta, Don and Eugene Charniak. 2000. Assigning function tags to parsed text. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the American Association for Computational Linguistics (NAACL-2000)*, pages 234–240,Seattle, WA.

Boonthum, Chutima, Shunichi Toida, and Irwin B. Levinstein. 2006. Preposition senses: Generalized disambiguation model. In *Proceedings of the Seventh International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2006)*, pages 196–207, Mexico City.

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

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

Carreras, Xavier and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Natural Language Learning (CoNLL-2004)*, pages 89–97, Boston, MA.

Carreras, Xavier and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, MI.

Cassidy, Patrick J. 2000. An investigation of the semantic relations in the Roget's Thesaurus: Preliminary results. In *Proceedings of the First International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2000)*, pages 181–204, Mexico City.

Cruse, David A. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

Edmonds, Phil and Scott Cotton, editors. 2001. *Proceedings of the Senseval 2 Workshop*. Association for Computational Linguistics, Toulouse.

Fillmore, Charles. 1968. The case for case. In Emmon Bach and Rovert T. Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart and Winston, New York, pages 1–88.

Fillmore, Charles J., 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*, pages 3–25, Hong Kong.

Frawley, William. 1992. *Linguistic Semantics*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Grozea, Cristian. 2004. Finding optimal parameter settings for high performance word sense disambiguation. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 125–128, Barcelona.

Kilgarriff, Adam. 1998. Senseval: An exercise in evaluating word sense disambiguation programs. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC '98)*, pages 581–588, Granada.

Koomen, Peter, Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2005. Generalized inference with multiple semantic role labeling systems. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL)*, pages 181–184, Ann Arbor, MI.

Lehmann, Fritz. 1996. Big posets of participatings and thematic roles. In Peter W. Eklund, Gerard Ellis, and Graham Mann, editors, *Conceptual Structures: Knowledge Representation as Interlingua*, Springer-Verlag, Berlin, pages 50–74.

Lenat, Douglas B. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Litkowski, Kenneth C. 2002. Digraph analysis of dictionary preposition definitions. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 9–16, Philadelphia, PA.

Litkowski, Kenneth C. 2004. Senseval-3 task: Automatic labeling of semantic roles. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 9–12, Barcelona.

Litkowski, Kenneth C. and Orin Hargraves. 2006. Coverage and inheritance in The Preposition Project. In *Third ACL-SIGSEM Workshop on Prepositions*, pages 37–44, Trento.

Litkowski, Kenneth C. and Orin Hargraves. 2007. SemEval-2007 task 06: Word-sense disambiguation of prepositions. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 24–29, Prague.

Liu, Rey-Long and Von-Wun Soo. 1993. An empirical study on thematic knowledge acquisition based on syntactic clues and heuristics. In *Proceedings of the 31st Annual Meeting of the Association for Computational*

Linguistics (ACL-93), pages 243–250, Columbus, OH.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcus, Mitchell, 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 *Proceedings of the ARPA Human Language Technology Workshop*, pages 110–115, Plainsboro, NJ.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mihalcea, Rada. 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)*, pages 1–7, Taiwan.

Mihalcea, Rada, Timothy Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4): Special Issue on WordNet.

Miller, Katherine. 1998. Modifiers in WordNet. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, pages 47–67.

Mohit, Behrang and Srini Narayanan. 2003. Semantic extraction with wide-coverage lexical resources. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 64–66, Edmonton.

O'Hara, Thomas P. 2005. *Empirical acquisition of conceptual distinctions via dictionary definitions*. Ph.D. thesis, Department of Computer Science, New Mexico State University.

O'Hara, Tom, Rebecca Bruce, Jeff Donner, and Janyce Wiebe. 2004. Class-based collocations for word-sense disambiguation. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 199–202, Barcelona.

O'Hara, Tom and Janyce Wiebe. 2003. Classifying functional relations in Factotum via WordNet hypernym associations. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003)*, pages 347–359, Mexico City.

OpenCyc. 2002. OpenCyc release 0.6b. Available at `www.opencyc.org`.

Palmer, Martha, Dan Gildea, and Paul Kingsbury. 2005. The Proposition Bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–106.

Palmer, Martha Stone. 1990. *Semantic Processing for Finite Domains*. Cambridge University Press, Cambridge.

Pradhan, Sameer, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2005. Support vector learning for semantic argument classification. *Machine Learning*, 60(1–3):11–39.

Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Scott, Sam and Stan Matwin. 1998. Text classification using WordNet hypernyms. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, pages 38–44, Montreal.

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

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

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

Srihari, Rohini, Cheng Niu, and Wei Li. 2001. A hybrid approach for named entity and sub-type tagging. In *Proceedings of the 6th Applied Natural Language Processing Conference*, pages 247–254, Seattle.

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

Wilks, Yorick, Brian M. Slator, and Louise Guthrie. 1996. *Electric Words*. MIT Press, Cambridge, MA.

Witten, Ian H. and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA.

Ye, Patrick and Timothy Baldwin. 2006. Semantic role labeling of prepositional phrases. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(3):228–244.

Ye, Patrick and Timothy Baldwin. 2007. MELB-YB: Preposition sense disambiguation using rich semantic features. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 241–244, Prague.