Inferential Commonsense Knowledge from Text

by

Jonathan Michael Gordon

Submitted in Partial Fulfillment of the
Requirements for the Degree
Doctor of Philosophy

Supervised by
Professor Lenhart Karl Schubert

Department of Computer Science
Arts, Sciences and Engineering
Edmund A. Hajim School of Engineering and Applied Sciences

University of Rochester
Rochester, New York

2014

To my parents.

Biographical Sketch

Jonathan Gordon was born in Wayne, NJ, in 1985. He attended Vassar College, where he was advised by Nancy Ide. At Vassar, he worked on the annotation the American National Corpus and, with Luke Hunsberger, performed research on learning models of musical chord progressions. He graduated with a Bachelor of Arts degree in Computer Science in 2007 and matriculated at the University of Rochester. In 2008, he was a visiting research assistant at the Institute for Creative Technologies at the University of Southern California, where he worked with H. Chad Lane on the automated analysis of student essays for an intelligent tutoring system. He was awarded the Master of Science degree in 2009. At Rochester, he pursued his doctoral research in artificial intelligence under the direction of Lenhart Schubert.

The following publications were a result of work conducted during doctoral study:

J. M. Gordon, B. Van Durme, and L. K. Schubert. Weblogs as a source for extracting general world knowledge. In Y. Gil and N. F. Noy, editors, *Proceedings of the Fifth International Conference on Knowledge Capture* (K-CAP), pages 185–6, Redondo Beach, CA, Sept. 2009. Association for Computing Machinery

———. Learning from the Web: Extracting general world knowledge from noisy text. In V. Nastase, R. Navigli, and F. Wu, editors, *Proceedings of the AAAI Workshop on Collaboratively-built Knowledge Sources and Artificial Intelligence*, pages 8–13, Atlanta, GA, July 2010b. AAAI Press

———. Evaluation of commonsense knowledge with Mechanical Turk. In C. Callison-Burch and M. Dredze, editors, *Proceedings of the NAACL Workshop on Creating*

Speech and Language Data with Amazon's Mechanical Turk, pages 159–62, Los Angeles, CA, June 2010a. Association for Computational Linguistics

J. M. Gordon and L. K. Schubert. Quantificational sharpening of commonsense knowledge. In C. Havasi, D. B. Lenat, and B. Van Durme, editors, *Proceedings of the AAAI Fall Symposium on Commonsense Knowledge* (*CSK*), Arlington, VA, Nov. 2010. AAAI Press

———. Discovering commonsense entailment rules implicit in sentences. In S. Pado and S. Thater, editors, *Proceedings of the EMNLP Workshop on Textual Entailment* (*TextInfer*), Edinburgh, Scotland, July 2011. Association for Computational Linguistics

———. Using textual patterns to learn expected event frequencies. In J. Fan, R. Hoffman, A. Kalyanpur, S. Riedel, F. M. Suchanek, and P. P. Talukdar, editors, *Proceedings of the NAACL Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction* (*AKBC-WEKEX*), pages 122–7, Montréal, Quebec, Canada, June 2012. Association for Computational Linguistics

———. WordNet hierarchy axiomatization and the mass–count distinction. In D. A. Evans, M. van der Schaar, and P. Sheu, editors, *Proceedings of the Seventh International Conference on Semantic Computing* (*ICSC*), Irvine, CA, Sept. 2013. IEEE

J. M. Gordon and B. Van Durme. Reporting bias and knowledge acquisition. In F. M. Suchanek, S. Riedel, S. Singh, and P. P. Talukdar, editors, *Proceedings of the Workshop on Automated Knowledge Base Construction* (*AKBC*), pages 23–30, San Francisco, CA, Oct. 2013. Association for Computing Machinery

Acknowledgments

> By the toil of others we are led into the preserve of things which have been brought from darkness into light.
>
> LUCIUS SENECA, *On the Shortness of Life*

This dissertation would not have been possible without many others, including

— my parents, for a life-long interest in what we can learn by reading;

— Molly, for her love and patience as I read and learned;

— the faculty and staff of the Department of Computer Science, for giving me the time, space, and support to pursue my research;

— my doctoral committee, James Allen, Gregory Carlson, and Daniel Gildea, for their instruction and their questions;

— my research collaborators, including Benjamin Van Durme, Fabrizio Morbini, and Adam Purtee, for sharing their ideas and their work;

— my advisor, Lenhart Schubert, for entirely too much of his time and knowledge.

Abstract

To enable human-level artificial intelligence, machines must have access to the same kind of commonsense knowledge about the world that people have. The best source of such knowledge is text – learning by reading. Implicit in linguistic discourse is information about what people assume to be possible or expect to happen. From these references, I obtain an extensive collection of semantically underspecified 'factoids' – simple predications and conditional rules. Using lexical-semantic resources and corpus frequencies, these factoids are generalized and partially disambiguated to form a collection of reasonable commonsense knowledge. Together with lexical axioms from the interpretation of WordNet, these probabilistic logical inference rules allow a reasoner to draw conclusions about everyday situations as might be encountered while reading a story or conversing with a person.

Contributors and Funding Sources

This work was supervised by a dissertation committee consisting of Professors Lenhart Schubert (advisor), James Allen, and Daniel Gildea of the Department of Computer Science and Professor Gregory Carlson of the Departments of Linguistics and Philosophy. Work in Chapter 3 and Appendix A was completed with additional guidance from Benjamin Van Durme, while the discussion and measurement of reporting bias in Chapter 4 is an expansion on points raised by Van Durme (2010) and benefitted from feedback by Lenhart Schubert, Peter Clark, and Doug Downey. All other work conducted for the dissertation was completed by the student independently.

This material is based upon work supported by NSF awards IIS-0535105 – *Knowledge Representation and Reasoning Mechanisms for Explicitly Self-Award Communicative Agents* (2006–9), IIS-0916599 – *General Knowledge Bootstrapping from Text* (2009–13), and IIS-1016735 – *Adapting a Natural Logic Reasoning Platform to the Task of Entailment Inference* (2010–12); DARPA N00014-11-C-0474 through subcontract from the Friedland Group; and a gift from Bosch Research and Technology. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of above named organizations.

Contents

List of Tables

List of Figures

1 Introduction

> To think is to forget differences, generalize, make abstractions.
>
> JORGE LUIS BORGES, 'Funes the Memorious', 1942

Artificial intelligence is concerned with the understanding of intelligence and the creation of intelligent artifacts, with the study of each informing the other. While impressive advances have been made at many subproblems of AI, these efforts often display a lack of the commonsense reasoning characteristic of human intelligence. For instance, when the Watson system competed at Jeopardy, a clue asked for 'the anatomical oddity of US gymnast George Eyser'. Watson responded, 'What is a leg?' This was good as a shallow match, but it failed to say that Eyser was *missing* the leg. The project head explained, 'The computer wouldn't know that a missing leg is odder than anything else' (Hamm, 2011).

To enable human-level artificial intelligence, it seems machines need access to the same kind of commonsense knowledge about the world that people have. This *knowledge acquisition bottleneck* is apparent not only in question-answering, but in other hard problems such as *natural language understanding* (Allen, 1994; Schubert, 2014). An intelligent agent should draw very different conclusions, for instance, when told 'Sarah had a pomegranate' and 'Sarah had a baby'. The need for accurate language understanding and commonsense reasoning ranges from mobile phone applications like Siri to counterterrorism intelligence efforts.

1.1 Commonsense Knowledge

Commonsense knowledge is the most fundamental and general knowledge about the world, shared by most people. No bright line distinguishes common sense from other varieties of general world knowledge or even from domain knowledge, making this a somewhat amorphous goal. Nonetheless, some of the varieties of commonsense knowledge an intelligent agent requires have been identified, *e.g.*, by McCarthy (1990):

> Common-sense knowledge includes the basic facts about events (including actions) and their effects, facts about knowledge and how it is obtained, facts about beliefs and desires. It also includes the basic facts about material objects and their properties.

Commonsense knowledge tends to be about *kinds* (*Countries have capital cities*) rather than individuals (*Edinburgh is the capital of Scotland*). It tends to be enduringly true (*People enjoy listening to music*) rather than true at the moment (*Barack Obama is president of the United States*). And it tends to be knowledge applicable to daily life (*Students attend classes*) rather than less familiar contexts (*A cloning vector replicates within a living cell*). As noted by Schubert & Hwang (1990), commonsense knowledge includes both predictive knowledge (*When an animal sees food, it may want to eat it*) and explanatory knowledge (*If an animal wants to eat some food it is probably hungry*). And we are especially interested in acquiring *causal* knowledge because it is pervasive in the understanding of narrative texts and a crucial part of our commonsense understanding of the world, including how people interact.

In this dissertation, I write of artificial agents acquiring or learning knowledge by reading text. Such claims are problematic philosophically: What basis is there for inductive reasoning, *i.e.*, for our generalization from the specifics being read to broad claims about the world? If a text is inaccurate or it is misunderstood by the system, can it be said to have learned or to know anything, even if the item of knowledge is true? If the result of this reading is inaccurate, is it *knowledge*? McCarthy & Hayes (1969) noted that

while it is important for researchers in artificial intelligence to consider these questions, many philosophical concerns become irrelevant when we undertake to create an agent. Doing so requires us to make simplifying assumptions, such as the existence of a physical world, our ability to obtain information about this world, and the correctness of our commonsense view of it. McCarthy wryly rebutted epistemological objections to AI in the dialogue at the end of 'Programs with Common Sense' (1959):

> Whenever we program a computer to learn from experience we build into the programme a sort of epistemology. It might be argued that this epistemology should be made explicit before one writes the programme, but epistemology is in a foggier state than computer programming even in the present half-baked state of the latter. I hope that once we have succeeded in making computer programs reason about the world, we will be able to reformulate epistemology as a branch of applied mathematics no more mysterious or controversial than physics.

1.2 Knowledge Representation

A precondition to learning and reasoning is a suitable way to represent knowledge. While we lack easy access – introspective or scientific – to human mental representation, we have language. Natural language is the universal human means of encoding and communicating knowledge (as well as other mental contents such as intentions), so it must be adequate to express our common sense. This section considers directly using language for knowledge representation and reasoning (KRR) and then a spectrum of more formal representations.

Natural Language as Knowledge Representation Abandoning formalisms and using natural language (NL) for knowledge representation and reasoning has appealed to many researchers. For instance, Singh (2002) argued for the use of NL in the Open Mind project's knowledge acquisition (see §2.2) and presented the Reform system for performing

forward inference over simple English sentences. Reform abstracted from example sentences to learn inferential patterns that could be applied to shallow syntactic parses to produce new sentences as conclusions. For instance, from sentences like

(A door) is a portal.
(Bob) opens (the door).
⇒ (Bob) can go through (the door).

Reform generalizes to the syntactic inference rule

(S (NP ?x) (VP is (NP ?y)))
(S (NP ?z) (VP opens (NP ?x)))
⇒ (S (NP ?z) (VP can go (PP through (NP ?x))))

Reform also used paraphrase (*i.e.*, bidirectional entailment) rules to convert between similar ideas, such as 'Joseph likes to drink wine' and 'If Joseph drinks wine, then he will enjoy it.' Other rules handled splitting and merging sentences, taxonomic inference ('Cats are mammals', 'Mammals are animals', therefore 'Cats are animals'), and inference about effects ('Pushing a door will open the door', 'Bob pushed the door', therefore 'Bob opened the door'). While this is an interesting idea, learning NL inference rules in this way is risky. The example above (his) would allow one to reason:

(A box) is a container.
(Bob) opens (the box).
⇒ (Bob) can go through (the box).

More recently, MacCartney & Manning (2008) presented Natural Logic (NLog), an approach to reason over syntactically – but not semantically – analysed language with a focus on the implicative and factive polarity of words. In Natural Logic, negation, monotonicity, implicatures, and lexical relations are modeled as part of a sentence's asserted content and treated through a projection mechanism. NatLog performs natural language inference (NLI) in five stages: linguistic analysis, alignment, lexical entailment

classification, entailment projection, and entailment composition. NLog was augmented by Clausen & Manning (2009) with the Karttunen (1973) model for lexically triggered presuppositions, such as the factive verb *know* in 'Bush knew that Gore won the election' presupposes that Gore won the election.

While natural language and Natural Logic can support simple knowledge representation and reasoning, a reasoner that is adequate for the purposes of artificial intelligence will need to address the range of phenomena that motivate a logical formalism, *e.g.*, quantifier scope, sentence embeddings, polarity, factivity, implicativity and exclusion, temporal and event relations among others. Therefore, it is desirable to use a representation that allows us to treat such phenomena *explicitly* and *consistently*. More formal representations can tolerate – or suffer from – some of the ambiguities of language, such as different, potentially overlapping senses of words, but in other ways they let us commit to a more precise meaning.

Logic for Common Sense Commonsense knowledge has been included in discussions of formal logic since Aristotle, though some philosophers, notably Wittgenstein, have argued that commonsense knowledge cannot be formalized or that logic is inappropriate for its representation.[1] Certainly commonsense knowledge poses representational problems that require more than traditional propositional or first-order logic (FOL). This section briefly traces developments in semantics, logic, and artificial intelligence that lead to the use of Episodic Logic in this dissertation.

The use of logic to express a program's knowledge and how it should reason was first proposed by John McCarthy for his *advice taker* (McCarthy, 1959). This ambitious paper led to decades of work on the appropriate use of logic to represent knowledge for

1 Wittgenstein (1953) wrote, 'You say: the point isn't the word, but its meaning, and you think of the meaning as a thing of the same kind as the word, though also different from the word. Here the word, there the meaning. The money, and the cow that you can buy with it. (But contrast: money, and its use.)' These and other remarks from Wittgenstein's later writing have been cited by Christopher Manning in arguing for Natural Logic rather than a more formal representation.

reasoning in AI. Motivated by problems in planning, McCarthy & Hayes (1969) presented the situation calculus, where a *situation* is the – real or hypothetical – state of the universe for a particular instant, incompletely expressed. In this representation, a *propositional fluent* like raining(x, s) is a function that maps situations (s) and other arguments such as the location x to true or false. A *situational fluent* maps situations to other situations, *e.g.*, result(p, σ, s) – the result of person p performing action σ in situation s.

In representing knowledge about the real world (rather than the mathematical world or the simple artificial worlds of early planning), it is necessary to express knowledge that is uncertain. Modal logic was created by Lewis (1918) to distinguish between those propositions that are *necessarily* true and those that are *contingently* true – *i.e.*, could be false. McCarthy & Hayes (1969) suggested the use of modal operators *normally*, *consistent*, and *probably* to deal with uncertain inference. They warn against attaching numeric probabilities to all logical statements as it is unclear how to do so for quantified statements 'in a way that corresponds to the amount of conviction people have'.

Nonomonotonic reasoning (NMR) allows a reasoner to draw conclusions that may be withdrawn in light of new information. For instance, the use of default rules reflects the human communication convention of expressing knowledge only if a person would not assume it. (The implications of this idea for knowledge extraction are considered in Chapter 4.) For a similar reason, when an agent learns rules about the world, these will necessarily overgeneralize. Although every object is in some way abnormal, we want to assume it is otherwise normal, in the absence of other information. As McCarthy (1986) wrote:

> Both common sense physics and common sense psychology use nonmonotonic rules. An object will continue in a straight line *if nothing interferes with it*. A person will eat when hungry *unless something prevents it*. Such rules are open ended about what might prevent the expected behavior, and this is required, because we are always encountering unexpected phenomena that modify the operation of our rules. [Emphasis mine.]

To express knowledge for NMR, McCarthy (1980, 1986) presented *circumscription*. In this approach, default rules use a single abnormality predicate *ab* taking as its argument an abstract entity corresponding to an aspect of the entities involved. Consistent with McCarthy's earlier conviction, circumscription is intended to be usable even when we do not have access to numeric probabilities about our knowledge. The classic example that 'birds fly' would be rendered

$$\forall x \,.\, \text{bird}(x) \wedge \neg\text{ab}(\text{aspect1}(x)) \Rightarrow \text{flies}(x)$$

To avoid claiming that penguins fly, we say they are abnormal in aspect 1:

$$\forall x \,.\, \text{penguin}(x) \Rightarrow \text{ab}(\text{aspect1}(x))$$

Pearl (1995) notes that in nonmonotonic logic, rules 'are usually interpreted as conversational conventions, as opposed to descriptions of empirical reality…' We don't, *e.g.*, require any statistical information about how many birds actually fly.

NMR would capture generalizations such as that 'Most (pet) dogs are friendly' with a rule of the type 'if x is a dog, you can conclude that x is friendly, unless you can prove otherwise'. However, in general this is neither effective (as provability is undecidable even in FOL) nor usable as a premise allowing us to infer, say, that many dogs are friendly (given that there are very many dogs). Nor is it easily adaptable to other nonclassical quantifiers, such as *many* or *occasionally*, *e.g.*, in 'Occasionally, a tree is struck by lightning in a thunderstorm'. Such quantified facts are important in language, commonsense reasoning, and life – and they are the kind of knowledge sought in §5.2.

Davidsonian Event Semantics A central issue in representing commonsense knowledge is the handling of events. Davidson (1967) gave a treatment of action sentences where the action verb has an event argument that is not explicit in language, *e.g.*,

Brutus killed Cæsar.

$$\exists e \,.\, \text{killed}(\text{Brutus}, \text{Cæsar}, e)$$

which can be read 'There is an event e such that e is the killing of Cæsar by Brutus'. This analysis allows the explicit temporal relation of events as in Davidson's example

Earwicker slept before Shem kicked Shaun.

$\exists e_1 \,.\, \text{slept}(\text{Earwicker}, e_1) \wedge$

$\quad \exists e_2 \,.\, \text{kicked}(\text{Shem}, \text{Shaun}, e_2) \wedge \text{before}(e_1, e_2)$

A limitation of Davidson's method of attaching event variables to atomic predicates is that it cannot represent sentences that describe events with quantifiers or negation, *e.g.*, 'Napoleon did not greet every general. This disappointed them.'

Hobbsian Logic Hobbs (1985) proposed a first-order, non-intensional logic where an English sentence is reduced to a conjunction of atomic predicates, and all variables are existentially quantified with the widest possible scope.[2] He follows Davidson in giving predicates an extra event argument, but doesn't limit this to action predicates. To allow an event argument for a standard eventless predicate, Hobbs also introduced the nominalization operator $'$:[3]

$$\forall x_1, \ldots, x_n \,.\, \text{p}(x_1, \ldots, x_n) \equiv \exists e \,.\, \text{Rexist}(e) \wedge \text{p}'(e, x_1, \ldots, x_n)$$

The *Rexist* predicate indicates that an eventuality exists not just in the Platonic universe of possible individuals but also in reality. For instance,

Brutus kills Cæsar.

kill(Brutus, Cæsar)

Rexist(E) $\wedge$ kill$'$(E, Brutus, Cæsar)

Brutus wants to kill Cæsar.

Rexist(E_1) $\wedge$ want$'$(E_1, Brutus, E_2) $\wedge$ kill$'$(E_2, Brutus, Cæsar)

2 He writes, 'Much of the complexity of English syntax, *e.g.*, the division of predicates into nouns, adjectives, verbs, adverbs, and prepositions reflects a conceptual scheme that is better captured in the axioms than in the syntax of our formal language…Morphemes introduce predications, and that's all.' (Hobbs, 2013)

3 Although so-called, it is not really an operator; there are simply two parallel sets of systematically named predicates (Hobbs, 2013).

The second example does not assert that the wanted event (E_2) actually exists, only the event of wanting it (E_1).

For Hobbs, there is no need for functions, predicate modifiers, nested quantifiers, disjunctions, negations, or modal/intensional operators. Instead, he relies on the nominalization operator and predicates applied to events, with quantified statements (*e.g.*, 'Most men work') taken as claims about typical elements of sets.

Neo-Davidsonian Event Semantics The Davidsonian representation was updated by Parsons (1990) using semantic roles associated with event variables to specify the various participants and properties of an event. There's no consensus in the choice of general semantic roles, but a possible neo-Davidsonian treatment is:

> Brutus stabbed Cæsar with a knife.
> $\exists e$. stabbing(e) $\wedge$ before(e, Now1) $\wedge$ agent(e, Brutus) $\wedge$ patient(e, Cæsar) $\wedge$ instrument(e, Brutus's-knife)

The slot-like use of thematic roles makes neo-Davidsonian representations compatible with frame-based knowledge, discussed later in this section.

Reichenbach Reichenbach (1947) proposed an analysis of sentences and events that anticipated the later work of situationists like Barwise and Perry. For Reichenbach, functions took physical objects and space-time locations as arguments. He introduced the $[]^*$ function to map sentences to such functions, *e.g.*,

> Sherlock met Moriarty at Reichenbach Falls[4] on Tuesday at noon.
> $(\exists v)$[meet(Sherlock, Moriarty, Reichenbach-Falls, Tuesday-12:00)]$^*(v)$

This means that the sentence 'Sherlock…' describes the fact v (where facts and events are conflated). While Davidson's account was limited to action sentences, Reichenbach's is general; his fact function can take compound sentences with quantifiers.

4 Sorry.

Situation Semantics To account for sentences about perception and attitudes, Barwise & Perry (1983) suggested that sentences refer not to truth values but to situations. They took meaning to be a relation between the situation in which a sentence is uttered, a connective situation, and a described situation. This approach was broadly influential but later foundered on issues such as the representation of negation.

Frames, Scripts, and Microtheories Work in artificial intelligence has involved the exploration of representations that provide unified bundles of knowledge rather than collections of knowledge fragments. Notably, Minsky (1974) introduced *frames* – data structures for representing stereotyped situations, including both knowledge and procedural information. Frames are hierarchical – inheriting properties from those that subsume them – and are intended to model expectations through the use of default assignments of the frame's *terminals* (slots). A *frame-system* could be used to represent actions or cause–effect relations by including different frames that share the same terminals. Frames have the advantage of coherently collecting knowledge about an entity and its attributes, parts, or participants.

Schank (1975) agreed that 'In order to build a real [natural language] understanding system it will be necessary to organize the knowledge that facilitates understanding. We view the process of understanding as the fitting in of new information into a previously organized view of the world.' Taking frames as a more general class of representations, he argued for the necessity of *script* and *plan* knowledge for understanding stories and actions respectively. Schank defined a script as a predetermined causal chain describing the normal sequence of things for a familiar situation (*e.g.*, eating at a restaurant), from a particular perspective (*e.g.*, that of a customer, a waiter, or a maître d'). These scripts can include essential parts and also alternatives. Plans are not fixed, allowing us to deal with new situations based on our goals and knowledge of the preconditions and effects of actions. More recently, the Cyc project (Lenat, 1995) has organized its knowledge base into explicit, hierarchical contexts in which the knowledge applies, including situations (*e.g.*,

a wedding, an office environment, shopping at a supermarket) reminiscent of Schankian scripts. (For more on Cyc, see §2.1.)

Organizing knowledge into coherent representations of common situations or sequences of actions and events is beyond the scope of this thesis. However, some of this knowledge is implicit in the knowledge learned about properties of classes and their generalizations and in extractions giving possible outcomes of actions, which can be chained. Future work can use clustering techniques to do so automatically along the lines of Chambers & Jurafsky (2008), described in §2.3.

1.3 Episodic Logic

Episodic Logic (Schubert & Hwang, 1990, 2000) is a situational logic designed to meet the representational requirements of natural language understanding and commonsense reasoning. Historically, Episodic Logic (EL) is a continuation of the work of Schubert & Pelletier (1982) to give a first-order logical form for English, inspired by Montague grammar. As EL will be used as the knowledge representation in the work that follows, in this section I give an overview of its expressivity (with reference to the work discussed in the previous section) and demonstrate notation.

In Episodic Logic, square brackets signify infix notation, used for most predication, *e.g.*, [Cæsar born-in Subura]. Round brackets signify prefix notation, used, *e.g.*, for functions and modification: (father-of Cæsar), (angrily (kill Cæsar)). Every infix sentence can be written equivalently in prefix notation. For instance, for a two-place predicate,

[B stab C] = [B (stab C)] = (stab C B) = ((stab C) B)

From Reichenbach and situation semantics, Episodic Logic adopts the use of sentences to describe episodes, which are limited pieces of reality. While the situation calculus of McCarthy & Hayes (1969) reasons about snapshots of the universe, in EL episodes can be temporally extended while their spatial extent and factual content may give them a more limited scope (Hwang & Schubert, 1993). Episodic variables are introduced

to make explicit the relationships among episodes – events, situations, circumstances, eventualities – which, in text, are often implicit and context-dependent. While Davidsonian episodic variables can only correspond to atomic formulas, in EL episodes can also involve quantification or negation, just as they do in English: 'Cæsar greeted every general. This [episode] made him tired.'

To say that a sentence *characterizes* an episode, EL uses the $**$ connective:

(∃*e*: [*e* before Now1]
    [(∀*x*: [*x* general.n]
        [Cæsar.name greet.v *x*]) $**$ *e*])

Here [*x* general.n] is a *quantifier restrictor*, which can be read 'every general'. When the scope of a quantifier is ambiguous, it can be written with angle brackets: [⟨an emperor.n⟩ greet.v ⟨every general.n⟩]. In linguistically derived knowledge, syntactic suffixes are typically used, as above, though EL does not require it.

A sentence characterizes an episode if it completely describes it, giving all the facts that are supported by it. EL also includes the weaker $*$, which connects a sentence to an episode it partially describes, *i.e.*, one in which it is true.[5] Any sentence that characterizes an episode necessarily also partially describes it. Note that episodes are distinct from actions (or activities), which specify an *agent*. The sentences

Brutus killed Cæsar (viciously).
Cæsar died (nobly).

can describe the same episode, but they describe different actions, which can be distinctly modified. Thus in EL an action is a pair of the agent *x* and the episode *e*, [*x* | *e*]:

(∃*e* [[[Brutus.name | *e*] (in-manner vicious.a)] ∧
    [[Cæsar.name | *e*] (in-manner noble.a)]])

5 Thus it is similar to Reichenbach's $[\varphi]^*(e)$ or Barwise's $e \vDash \varphi$.

Episodic Logic also allows for the reification (*i.e.*, nominalizing) of sentences and predicates, as in

Cæsar believes that Brutus is loyal.
[Cæsar.name believe.v (that [Brutus.name loyal.a])]

Blood is red.
[(k blood.n) red.a]

Here *k* is the kind-forming operator for nominals. After Carlson (1982), mass or abstract nominals and bare plurals in English are understood to refer to kinds. There is also *ke* to form kinds of events and *ka* to form kinds of actions. Schubert & Hwang (1990) give the examples:

For Mary to dance was rare.
[(ke [Mary dance]) rare]

To kiss Mary is fun.
[(ka (kiss Mary)) fun]

Another feature of Episodic Logic is the modification of predicates and sentences, *e.g.*,

Canada is very distant from Australia.
[Canada.name (very λx [x distant-from.a Australia.name])]

where *very* is a predicate modifier – a function that takes a predicate as its argument and returns a more restricted predicate. Other predicate modifiers can be formed using *adv-a*, which takes a one-place predicate over actions (agent–episode pairs), and *adv-m*, for manner adverbials:

[John.name ((adv-a (in-manner polite.a)) (greet.v Mary.name))]
[John.name ((adv-m sound.a) sleep.v)]

To form sentence-modifiers, EL provides *adv-e* for temporal or locative modification of an episode, and *adv-f* for a frequency-modifying adverbial:

((adv-e (during.p 15_March_44BCE.name)) [Brutus.name kill.v Cæsar.name])
((adv-f regular) [Senate.name meet.v])

For the purposes of this thesis, a central feature of Episodic Logic is its support for uncertainty and genericity, as in these sentences:

Cats meow.
A wolf is usually grey / Most wolves are grey.
If Allen wins an award, he will probably accept it.

Older work on EL, *e.g.*, Schubert & Hwang (1990) used a 'parameter mechanism' inspired by discourse representation theory (DRT) to allow episodes to persist between sentences, resulting in probabilistic general claims of the form:

A wolf is usually grey / Most wolves are grey.
$(\exists x\ [x\ \text{wolf.n}]) \Rightarrow_{.8} [x\ \text{grey.a}]$

Here the subscripted numbers are lower bounds on epistemic probabilities. Given the knowledge [W wolf]$^{.9}$, that paper's Rule Instantiation inference rule would give the conclusion [W grey.a]$^{.72}$.

In this work, such knowledge is treated either as the predication of a kind-level property (like 'extinct') of a reified kind or as the quantified predication of an instance-level property (like 'grey') over instances. This uses generalized quantifiers including *all-or-most*, *most*, and *many* that have associated numeric weights to enable probabilistic inference. (See §6.2 for details.)

In summary, Episodic Logic is designed to echo the expressive devices found in all human languages, including modification, reification, and vague generalized quantifiers. The use of such a representation is in tune with recent trends in textual inference, such as the growing use of Natural Logic (*e.g.*, MacCartney & Manning, 2008) or of logical forms that are essentially parse trees, perhaps with replacement of some subtrees by variables (Bar-Haim & Dagan, 2008). Indeed, the inference engine for EL, called Epilog, is

based on NLog-like monotonicity and entailment principles, along with natural deduction rules that (unlike NLog) allow for inferences based on multiple premises, including background world knowledge (see Schubert *et al.*, 2010).

1.4 Overview of the Thesis

In this thesis, I address the large-scale acquisition of world knowledge from text – learning by reading. Rather than seek a set of relations about specific individuals or target a restricted class of general relations, we look for all world knowledge that can be learned from a text. The result is a collection of symbolic logical formulas, which are automatically verbalized into English, *e.g.*, *An article may be published on a website* (Chapter 2). Using a variety of filtering techniques, we find that even 'noisy' text like that found on the Web yields factoids that human judges rate comparably to those learned from edited, informative sources like Wikipedia (Chapter 3). As we 'read' more sentences – hundreds of millions of them – the number of unique factoids grows logarithmically with the total number of factoids extracted.

Yet even in Web-scale text, commonsense knowledge is rarely stated. Consistent with Grice's maxim of quantity (Grice, 1975), the more universally people know something, the less likely they are to communicate it. Thus we mine the knowledge that is *implicit* in written language, abstracting from simple references ('John's brother': *A male may have a brother*) and from full sentences ('Germany elected a new president last year': *A country may elect a president*). Even so, it's a concern that people tend to discuss the unusual or unexpected. This *reporting bias* (Chapter 4) is mitigated by focusing on language indicating disconfirmed expectations (§5.1): If we read that 'the bomb was dropped but didn't explode', we learn that – more usually – *If a bomb is dropped, it may explode*.

While knowledge-extraction systems have learned quantificationally underspecified logical forms or simpler outputs such as tuples of normalized strings, these are unsuitable for use in a general reasoning engine like Epilog. Therefore, I create explicitly quantified, partially disambiguated axioms (§5.2): While a factoid might claim that *A person*

may have a head, this can be 'sharpened' to an axiom claiming *Every person has a head as a body part*. *A person may say something* is sharpened to *All or most persons at least occasionally say something*. This transformation depends on semantic categories (*e.g.*, having a body part *vs* a possession), semantic properties of predicates (*e.g.*, kind-level, individual-level, and repeatable or non-repeatable stage-level), and corpus frequencies (giving the strength of association between a subject and what's predicated of it). In later work (§ 5.3), I use textual patterns to learn the likely frequencies of regular, repeatable actions, *e.g.*, a person tends to sleep daily, but if they go to church they tend to do so weekly.

Chapter 6 demonstrates inference with the resulting knowledge, with generalized quantifiers like *all-or-most* giving probabilistic conclusions that are judged favorably compared with pseudo-inference performed using unsharpened factoids or the results of a recent information-extraction system. I close with a summary and some directions for future study (Chapter 7).

Work in knowledge extraction is intended to improve many applications, such as machine reading, question answering, and intelligent assistants. As a grand goal motivating this research, consider the creation of a broadly knowledgeable dialogue agent, as described by Schubert (2009). Such a system would require a general representation language with the expressiveness of natural language, such as Episodic Logic (§1.3); a general reasoner, such as Epilog (Schaeffer *et al.*, 1993; Morbini & Schubert, 2009); a semantic parser to generate appropriate logical representations of natural-language inputs; a dialogue manager; and – the focus of this dissertation – a large knowledge base of general world knowledge to support the processes of language understanding and reasoning.

While such a system is a strong motivation for work in knowledge extraction, it is not obvious that a dialogue agent, nor most of the other intended applications of commonsense knowledge, are necessarily good means of evaluating the knowledge itself. Traditionally, evaluation of knowledge-extraction systems has involved the authors or other experts rating a small sample of the output. In Appendix A, I demonstrate the use of un-

trained crowds to provide judgements that are well-correlated with those of experts. This allows for larger-scale evaluation or a means of human-filtering to create a high-quality 'core' knowledge base in the future.

And while this work acquires large amounts of knowledge from mining real-world text, existing semantic resources are also important as sources of commonsense knowledge. In Appendix B, I show the creation of lexical axioms, *e.g.*, *Gold is a noble metal* or *Every document is an amount of written material*, derived from WordNet's hypernym hierarchy. This work is based on the automated annotation of nominal word senses with relevant semantic properties, most crucially the mass–count distinction. The resulting axioms allow inference using knowledge stored at different levels of specificity.

2 Knowledge Acquisition

> We can only speculate that we are drawing on some basic human instinct to pass on our commonsense to our progeny.
>
> PUSH SINGH, 'The Public Acquisition of Commonsense Knowledge', 2002

There is a diversity of goals in knowledge acquisition, ranging from work finding uninterpreted text strings that may satisfy a specific relation (*e.g.*, hyponymy, paraphrase) to open-ended logical interpretation. Consequently, the approaches to acquiring knowledge also vary considerably. Although this survey is far from comprehensive, I describe several representative and interesting approaches before proceeding.

2.1 Knowledge Engineering

The most direct approach to acquiring the knowledge needed for reasoning is for researchers to write it down in the form they require. This work shares its approach with the traditional construction of resources for human use, such as dictionary and encyclopedia writing. It is distinguished from these endeavours both by its intended use and by its format; the knowledge in an encyclopedia or dictionary is encoded in natural language for use by people, while knowledge engineering produces resources in a directly machine-readable format.

Knowledge engineering (KE) guarantees high-quality results, but it is prohibitively slow and expensive. However, manual knowledge-engineering can be the best approach when knowledge is needed for a small domain or when there are inadequate resources

from which to learn. In many situations, KE can be supplemented with automatic extraction. Even when the need for accuracy is paramount, it may be more efficient to use a hybrid approach combining automatic extraction with manual verification or correction.

WordNet WordNet (Fellbaum, 1998) is a lexical-semantic database, which enumerates senses of nouns, verbs, adjectives, and adverbs, and organizes them into sets of synonyms (*synsets*) representing concepts. The synsets are linked by relations including *meronymy* (part-of) and *hypernymy* (is-a). WordNet's nominal hypernym hierarchy is often used for natural language understanding tasks or treated as an ontology for reasoning. I report problems of WordNet's knowledge for reasoning – and my solution to these – in Appendix B. WordNet synsets also include dictionary-style definitional glosses and examples which have been the target for work on acquiring formal lexical knowledge (*e.g.*, Clark *et al.*, 2008b).

YAGO Suchanek *et al.* (2007) created YAGO, a manually constructed ontology augmented with factual extractions connecting entities. From Wikipedia's division of articles into categories such as '1954 deaths' or 'Nobel laureates in Physics', the authors extracted instances of 14 predefined relations, *e.g.*, *AlbertEinstein hasWonPrize NobelPrize*. This resulted in over five million such facts, expressed in a variant of the Web Ontology Language (OWL). By focusing on the manually contributed knowledge expressed in Wikipedia's non-textual content, they avoid the difficulties of language understanding and favor precision over greater coverage. The authors estimate the contents of YAGO are 95 percent accurate.

FrameNet The FrameNet project (Fillmore & Baker, 2009) seeks to build a database of hundreds of frames (§1.2) with example sentences that support the analyses. FrameNet enumerates as *frame elements* common aspects or components of a frame, *e.g.*, for the Revenge frame, these include the Offender, InjuredParty, Avenger, Offense, and Punish-

ment. It also includes a list of words that evoke the frame, *e.g.*, *avenge*, *get even (with)*, *reprisal*, *revenge*, *vindictive*, *quid pro quo*, *etc.* Annotations of example sentences that use these trigger words show how the frame elements can be presented linguistically. Note that the individual words that evoke a frame can differ significantly in meaning, as for the Change of Phase frame, which includes *freeze* and *defrost* – these can be thought of as invoking their own frames that inherit from Change of Phase but with different before and after states. FrameNet has not included knowledge about negation or quantification.

Component Library To enable domain experts to build knowledge bases more quickly, Barker *et al.* (2001) created a hierarchy of manually constructed *components* – frame-like representations of common entities and events, with events including actions and the (relatively) static states produced by them. These basic components contain axioms describing how they interact with other components, allowing them to be composed into new concepts. The Component Library includes knowledge about the preconditions and postconditions of actions, such as that before an *enter* action, the object is in a state of being outside of an enclosure while after the *enter* it is in a state of being inside the enclosure, and that if the portal the object is passing through has a covering (*e.g.*, a door), it must be open. Unfortunately, the frame-based knowledge representation used in this project is not language-like.

Commonsense Psychological Axioms Hobbs & Gordon (2005) presented work to write a collection of axioms about they way people 'think that they think' about beliefs, plans, goal, *etc.* in Hobbs's flat FOL representation (described in §1.2). This work is a promising attempt to enumerate some of the most basic knowledge needed for reasoning that may largely lie outside the scope of what we can hope to learn from text.

Cyc The Cyc project (Lenat, 1995) is the largest knowledge-engineering effort designed to support artificial intelligence. Cyc was motivated as an effort to 'prime the pump' of automatic knowledge extraction – and AI generally – through the manual enumeration

of a million logical axioms by dozens of professional knowledge engineers. (As Curtis *et al.* (2009) report, Cyc's KB contains more than five million assertions.) From the perspective of this dissertation, Cyc is most interesting in its encoding of commonsense knowledge, including that *You have to be awake to eat*, *You can usually see people's noses, but not their hearts*, and *You cannot remember events that have not happened yet* (examples due to Lenat, 1995).

Matuszek *et al.* (2006) describe Cyc's knowledge base as dividing into an *upper ontology* for knowledge about mathematics and meta-knowledge about the classes that organize the lower levels, a *middle ontology* that stores general world knowledge, and a *lower ontology* that stores facts about particular instances, such as the names of politicians and the locations of events. While Cyc is the most prominent example of knowledge-engineering, its lower ontology has been expanded with the use of automatic knowledge extraction, as by Matuszek *et al.* (2005).

Cyc is notable for being logically precise about matters that can be very ambiguous in natural language. *E.g.*, CycL (Matuszek *et al.*, 2006) includes notation to handle the use–mention. For a standard predicate like *Dog*, *#$Dog* denotes the *term* 'dog'. This allows the natural expression of knowledge about the words used as predicates, such as when a term was coined. Another common ambiguity is that between (the abstract idea of) an authored work and the specific instantiations of it. *E.g.*, I might ask 'Have you read *Invisible Cities*?' referring only to the text, but I can also say '*Invisible Cities* is in the living room', referring to a particular physical copy. (A similar distinction applies to a play's text and its performances.) In Cyc, *PropositionalConceptualWork* denotes the class of abstract works that convey propositional content, whereas *InformationBearingThing* represents the collection of objects and events that might carry this information. This gives us separate predicates, *Book-CW* for a book as a conceptual work and *BookCopy* for the physical instantiation (Curtis *et al.*, 2009).

While Cyc knowledge bases have been used, *e.g.*, in the work of Forbus *et al.* (2009) toward learning-by-reading, they have not seen wide deployment in AI applications. This

is attributable in part to a general resistance to logical forms in contemporary AI research. *E.g.*, Havasi *et al.* (2007) write, 'To use Cyc for natural language tasks, one must translate text into CycL through a complex and difficult process, as natural language is ambiguous while CycL is logical and unambiguous.' In particular, Cyc's representational commitments mean it is not easily integrated with linguistically derived knowledge: Cyc collapses complex English constructions into atomic logical concepts, so, *e.g.*, the relation between *killing* and *dying* is expressed using *lastSubEvents*, *KillingByOrganism-Unique*, and *Dying* and requires the higher-order relation *relationAllExists*.[1]

2.2 Crowdsourced Knowledge Engineering

The Internet has facilitated the distributed construction of knowledge resources – most notably the user-created encyclopedia Wikipedia. Researchers have also attempted to acquire knowledge for machine use by forms of *crowdsourcing* – replacing professional knowledge engineers with large numbers of non-expert participants coordinated online. As Singh (2002) wrote, 'Every ordinary person has the common sense we want to give our machines', so why not ask them for it?

Open Mind and ConceptNet While knowledge engineers produce formal representations, crowdsourced workers are generally not capable or willing to do so. As such, efforts like the Open Mind Common Sense project (Singh, 2002) solicit knowledge in natural language. The project's response to this limitation has been, first, to argue for natural language as an appropriate knowledge representation for reasoning (as discussed in §1.2) and, second, to process the natural language responses they solicit into a new representation, ConceptNet.

ConceptNet (Singh, 2002; Liu & Singh, 2004; Havasi *et al.*, 2007) is a semantic network joining concept nodes by a small set of predefined relations, including temporal,

1 Example from Schubert (2014).

spatial, and causal ones. It is automatically constructed from the semi-structured sentences in the Open Mind Common Sense corpus by applying a set of rules that identify short English fragments for predicate relations and arguments. Open Mind has also solicited knowledge from users with activities that use template-based input. By asking users to enter knowledge into narrow fields, they focus contributions on the relations the researchers are interested in, *e.g.*,

> A *hammer* is for ___.
> The effect of *eating a sandwich* is ___.

The text fragments are turned into nodes by filtering punctuation, stop-words, stop-parts-of-speech, stemming, and then alphabetizing the results. This kind of normalization collapses text fragments with similar meanings, though it does so at the expense of merging some distinct concepts, *e.g.*, 'tap water' and 'water tap'.[2] Havasi *et al.* (2007) consider this the appropriate granularity for reasoning with language, even when it collapses phrases 'that are only related by accidents of orthography'. ConceptNet's nodes tend to represent verbs only in complete verb phrases like 'go to the store' rather than the bare verb 'go'. *E.g.*, in ConceptNet,

> You may be hurt if you get into an accident.
> EffectOf('get into accident', 'be hurt')

These effect relations can be chained to give significantly underspecified script-like knowledge. Liu & Singh (2004) give a possible chain connecting two verbal nodes in ConceptNet:

> 'buy food' ⇒ 'have food' ⇒ 'eat food' ⇒ 'feel full' ⇒ 'feel sleepy' ⇒ 'fall asleep'

The Open Mind website also allows participation in the filtering of knowledge by asking users to rate whether previous statements on a given topic are 'helpful, correct

2 Example due to James Allen.

knowledge' (Havasi *et al.*, 2007). Chklovski (2003) also allowed the Open Mind website to fill in missing but supposed knowledge by asking users to verify analogous claims. For instance, knowing that *newspapers* and *books* share properties such as *have pages*, if it is told that a book may be burned, it will ask a user whether a newspaper can also be burned.

The statements contributed to Open Mind vary significantly in quality, and the user ratings are an unreliable indicator of which statements contain accurate world knowledge. For instance, users sometimes seem to rate a statement based on whether they agree with its sentiment rather than whether it is appropriate world knowledge, giving *A friend in need is a friend indeed* the rather high score of 9.[3] The users contributing knowledge can also be imprecise about problems like the strength of assertions, as in *Sometimes having a haircut causes you to have shorter hair*, which has a score of 11. While this is true, a haircut *always* results in shorter hair.

Verbosity Another effort with an interesting approach to motivating contributions is Verbosity, 'a fun game with the property that common-sense facts are collected as a side effect of game play' (von Ahn *et al.*, 2006). Players are asked to fill in templates such as '__ is a kind of __' or '__ is typically near __', with game play consisting of one player trying to guess a word based on the information entered into these templates. Given the secret word 'computer', a player might fill in a template to tell the one who's guessing that 'It contains a keyboard'. While it is not clear that all or even most knowledge acquisition tasks can be readily turned into *games with a purpose*, when they can it provides a non-monetary incentive for participation.

3 The example is from a snapshot released in June 2008. There is no fixed range for Open Mind ratings as they consist of adding 1 whenever a statement is rated up and subtracting 1 whenever a statement is rated down. Thus a rating of 0 might indicate a sentence that has never been rated or one that is controversial and has been rated up and down many times.

2.3 Automatic Knowledge Extraction

Information extraction is an area of research that has seen considerable interest in recent years, due to the proliferation of information available in digital form and the demands of applications such as question-answering systems, which exceed the practical limits of manual enumeration. Information extraction focuses on the study of algorithms for automatically acquiring large collections of information with high accuracy. This can be augmented with a 'human in the loop' to maintain some of the benefits of knowledge engineering or crowdsourcing but on a larger scale and with a more systematic connection to natural language (see, *e.g.*, Hoffman *et al.*, 2009, and Appendix A).

While the goal of this thesis – the acquisition of commonsense knowledge suitable for reasoning – falls under the heading of information extraction, that term is most often applied to research that focuses on *fact extraction*. This is the problem of looking for information about specific individuals and events, such as *Alan Turing died in 1954* or *The capital of Bahrain is Manama*. I contrast this with the problem of *knowledge extraction*, which seeks to acquire more general claims. This will include some knowledge about specific individuals, but of a more general sort, *e.g.*, *The United States Congress may pass a law* or that *The Earth orbits the Sun*. While fact extraction is immediately useful for problems like basic question answering, we consider knowledge extraction necessary for *artificial intelligence*.

Reading Dictionaries It is natural to consider learning by reading intentionally informative sources, and much early work looked at the (partial) interpretation of dictionary definitions (surveyed by Ide & Véronis, 1994). However, dictionaries are resources for people who already have commonsense knowledge and are expected to consult definitions either for more obscure words/concepts or to obtain a technical definition of a concept they're already familiar with. As such, work in learning from dictionaries suffered from many problems, including circular definitions or larger conceptual loops and difficult, obtuse descriptions of common concepts. *E.g.*, the OED (Simpson, 2013)

defines the most common sense of 'house' as 'A building for human habitation, typically and historically one that is the ordinary place of residence of a family'. Most knowledge-extraction tools would have more success with the simple claim that 'A house is a building that a person lives in' or 'People live in houses'.

Hypernym Relations Hearst (1992) presented a method of template-driven extraction for the automatic extraction of *hypernymy* (or *is-a*) relations from text. As in WordNet, this relation holds between two terms *a* and *b* when English speakers accept sentences such as 'An *a* is a (kind of) *b*'. For instance, 'cat' is a hyponym of 'animal' and 'animal' is a hypernym of 'cat'. Hearst manually authored templates to match the expression of this relation in text, in constructions like 'such *b* as *a*'. (Later work, *e.g.*, Pantel & Pennacchiotti (2006), has learned these templates based on a number of seed examples.)

Examining the results of this extraction method, Hearst identified issues that appear in most work on knowledge extraction, including this dissertation: Knowledge may be found at an inappropriate level of specificity (in this case, the hyponym might be matched with a hypernym that is too high in the hierarchy). In other cases, matches may be dependent on textual context or a particular point of view rather than generally true.

Causal Extraction Girju (2003) presented work on extracting from text causal relations like *Earthquakes cause tidal waves*. She focused on event nominalizations linked in the general form 'NP_1 causal-verb NP_2', where the causal verb can be a phrase like 'set in motion', and the nominal arguments are restricted to those in 'causation classes' identified in WordNet so that, *e.g.*, 'the trail leads to…' does not generate a claim about a trail causing something. Causal knowledge is a central part of our commonsense understanding of the world, and the work presented in Chapter 5 acquires causal knowledge as part of a more general knowledge extraction process.

Verb–Verb Relations Chklovski & Pantel (2004) used Hearst-like manual lexico-syntactic patterns, sent to an Internet search engine, to find possible relations between verbs,

e.g., 'buy' may *happen-before* 'sell', and 'fight' may *enable* 'win'. One pattern given for finding enablement relations is '*x*-ed * by *y*-ing the', which might be instantiated 'obtained money by borrowing from…' or 'fixed the computer by plugging the…' They used the resulting co-occurrence counts to measure mutual information between pairs of verbs, and hence to assess the strengths of the relations. As with the chains of ConceptNet nodes, these results are underspecified with respect to any arguments. It's inadequate for inference to know that *Crashes cause injuries* without knowing what might be crashing – *e.g.*, cars, computers, stocks – and who or what it is that might be injured as a result.

Event Chains From news text, Chambers & Jurafsky (2008) induced *narrative event chains*. These similar to Schankian scripts, consisting of a partially ordered sets of events (predicate–argument pairs) involving a particular individual. After parsing and resolving coreference for a text, they count pairs of verbs that share coreferencing arguments and compute the pointwise mutual information (PMI) between the verb–argument pairs. Narrative event chains are created by clustering events' slots using their PMI scores and classifying events temporally. While their approach is similar to that of Chklovski & Pantel (2004) in using a distributional scoring metric, it differs in using references to a single *protagonist* as its indicator of relatedness.

Chambers & Jurafsky (2009) extended this work to learn *narrative schemas*, where semantic roles are found for the participants in these events, *e.g.*, arrested(*Police*, *Suspect*), with *Police* defined over specific words like {police, agent, authorities}. They merge verbs in distinct narrative chains into a single narrative schema, with the shared arguments across verbs allowing them to induce semantic roles. (This later work did not include decisions about the temporal ordering of events.)

Distributional Learning of Rules Zelig Harris's *distributional hypothesis* (Harris, 1985) states that words that occur in similar contexts have similar meanings. This idea was adapted by Lin & Pantel (2001) to sentence fragments as the *extended distributional hypothesis*, based on which, they:

1 Gather a large collection of dependency parses of sentences.
2 Identify the basic noun phrases in each sentence.
3 Collect all paths that connect similar nouns.

This gives an unsupervised method to learn from text semantically similar ideas, including rough paraphrases, entailments, and associated possibilities.The result, DIRT, includes approximately 12 million rules (with associated confidence values). Szpektor *et al.* (2007) give examples of DIRT's templates, with manual classification of incorrect rules and the entailment direction for correct rules:

Correct	*Incorrect*
x change $y \leftrightarrow x$ modify y	x change $y \approx x$ adopt y
x change $y \leftarrow x$ amend y	x change $y \approx x$ create y
x change $y \leftarrow x$ revise y	x change $y \approx x$ stick to y

While it has largely been assumed that work like DIRT will learn paraphrases, Szpektor *et al.* found that it mostly learned one-directional entailment rules.

Pantel *et al.* (2007) learned 'inferential selectional preferences' (ISPs) that constrain the arguments of DIRT inference rules to avoid certain implausible conclusions. However, these can still be too general. Clark & Harrison (2009a) gives the example rule 'If x shoots y then x injures y', which includes artifact#1 as a preference for y. This allows a system to conclude that 'Fred shoots the gun' implies that 'Fred injures the gun'.[4]

The KnowItAll Project KnowItAll (Etzioni *et al.*, 2004) is a system for domain-independent information extraction. To learn instances of a new property, the developer gives the system a set of examples, which it sends as queries to Internet search engines. From these results, it learns the frequency with which the property is expressed via generic Hearst

4 In that paper, Clark & Harrison blocked ISPs that were not supported by the extractions from their Knext-inspired DART tool. This resulted in an improvement at the use of such rules for recognizing textual entailment.

patterns including '*NP1* {,} such as *NPList2*' and '*NP1* is a *NP2*'. A naïve Bayes classifier is trained to judge whether new terms satisfy the property. KnowItAll then sends its stored contexts as queries in order to classify the terms that fill particular contextual slots in the returned queries.

This was succeeded by TextRunner (Banko *et al.*, 2007), which makes a single pass over a text corpus, extracting all relational tuples it finds, allowing it to avoid KnowItAll's dependence on manually selected relations with examples. As its creators consider syntactic parsing too slow for a scalable tool, TextRunner only tags words with their part of speech and identifies noun phrases with an NP chunker. It then forms a tuple for each pair of nearby NPs, with the intervening text identifying the relation. As a step toward abstraction, TextRunner drops adverbial and prepositional modifiers. The resulting tuples include the same knowledge with various numbers of arguments, *e.g.*, Van Durme & Schubert (2008) give the example

(the people, use, force)
(the people, use, force, to impose, a government)
(the people, use, force, to impose, a government, on, an economic base)

While TextRunner is primarily a fact extraction tool, this example shows that it can also identify some kinds of general world knowledge when explicitly stated.

TextRunner uses a Bayesian classifier trained on a small parsed corpus to label extractions as trustworthy or not, and a redundancy-based assessor assigns a probability to each of the trustworthy tuples based on a probabilistic model of redundancy in text. It does not attempt to convert the resulting tuples into a more formal representation. For the knowledge extraction community, TextRunner was most significant in introducing an emphasis on scaling extraction to run on large collections of text, as found on the Web.

While TextRunner collects tuples of information stated explicitly in text, Sherlock (Schoenmackers *et al.*, 2010) inferred first-order Horn-clause rules implicit in these results, *e.g.*,

IsHeadquarteredIn(Company, State) ⇐ IsBasedIn(Company, City) ∧ IsLocatedIn(City, State)

*ReturnTo(Writer, Place) ⇐ BornIn(Writer, City) ∧ CapitalOf(City, Place)

The second rule is unsound. A limitation of this approach is that, operating on the extractions of a factual IE system, they only learn rules involving the relations it discovers, which tend to be about simple attributes like locations or roles rather than consequences or reasons.

Fader *et al.* (2011) attacked the problem of incoherent and uninformative TextRunner extractions by introducing syntactic and lexical constraints at the expense of limiting output to binary relations expressed by verbs. (The possible use of these results for inference is evaluated in §6.3.2 as a baseline for the results of this dissertation.) KrakeN (Akbik & Löser, 2012) is designed to surpass ReVerb by extracting relations with more than two arguments. For instance, from the sentence 'Elvis moved to Memphis in 1948', ReVerb learns only MovedTo(Elvis,Memphis), while KrakeN would capture the temporal argument as well. KrakeN has a lower extraction rate than ReVerb, but its output was evaluated as higher precision, and more extractions were rated as 'complete', *i.e.*, having all necessary arguments.

Banko & Etzioni (2007) considered the problem of building a 'lifelong learning agent', Alice, that takes the output of TextRunner and creates *domain theories* to more compactly represent related knowledge, *e.g.*,

Instance: *Orange* is an instance of *Fruit*

Attribute/relation: *Fruit* is something that *Grows*

Relationship: GrowIn(*Fruit*, *Location*)

General proposition: Provide(*Fruit*, *Vitamin*)

These domain theories are updated with prior learned knowledge guiding the system's decision about what to try to learn next. General propositions are found by a process of abstraction, with the proposition above being deduced from the tuples

(oranges, provide, vitamin c)

(bananas, provide, a source of B vitamins)

(an avocado, provides, niacin)

This abstraction presents the problem of ensuring that a general proposition is of the appropriate level of generality, covering all the related instances but not overgeneralizing. Should *fruit* provide *vitamin* or *substance*? They try to find the lowest point in their concept hierarchy that describes a relation, relying on a clustering approach.

NELL Carlson *et al.* (2010) presented work on an agent that improves its ability to learn category instances, *e.g.*, London is a city, and pre-specified semantic relations, *e.g.*, hasOfficeIn(BBC, London), from free-form text and from semi-structured data such as tables or lists. NELL begins with 10–15 labeled seed instances for each of its categories and five initial Hearst patterns. It trains a model and then uses that model to label more data – semi-supervised bootstrap learning. While the varieties of knowledge learned by NELL are found by previous systems with high accuracy, its emphasis on bootstrapping is an important direction for future work in knowledge extraction.

Forbus On the other side of the breadth–depth divide is the work of Ken Forbus and his group on learning by reading. The Learning Reader system (Forbus *et al.*, 2007) extracts knowledge from short stories written in simplified English. It uses a Direct Memory Access Parsing (DMAP) model of natural-language understanding as the recognition of concepts based on the phrasal patterns of their expression in text – in this case using mappings from 30,000 phrasal patterns to Cyc concepts. Learning Reader includes a model of *rumination* – asking itself questions in order to assimilate new knowledge, *e.g.*, for each event trying to answer the standard 'who', 'what', 'when', 'where', and 'why' questions. In an experiment, this was found to boost the system's ability to answer questions, increasing coverage from 37% to 50% with only a small drop in accuracy.

2.4 Episodic Logic Interpretation and Abstraction

Previous work is unsatisfactory in several ways, including:

1. Many information-extraction efforts only look for a small set of predefined relations.
2. Statistical or pattern-matching techniques often require great redundancy of information. Even on the vastness of the Web, important commonsense knowledge may be stated quite rarely. In fact, the more basic the knowledge, the less likely it is to be mentioned. (See Chapter 4.)
3. Systems produce simple representations such as tuples of text strings or binary relations between concepts, which cannot represent the variety of knowledge people know, and are unsuitable for drawing conclusions by inference.

Machines meant for interacting with people using natural language require the ability to represent and to reason with the full range of complex phenomena seen in natural language. As presented in § 1.3, Episodic Logic is designed to support these linguistic phenomena, both as a knowledge representation and as a semantic representation. In this section, I present a synopsis of previous work in the interpretation of the explicit content of natural language into Episodic Logic and its abstraction to form collections of world knowledge.

Early Episodic Logic Interpretation In Montague grammar (Thomason, 1974), for every syntactic composition there is an analogous rule of semantic composition. Inspired by this idea, Schubert & Pelletier (1982) presented the use of semantic interpretation rules to generate *immediate logical forms*[5] – representations that do not attempt to capture the full meaning of an utterance but permit ambiguity in quantifier scope, the identity of referents, the sense of predicates, *etc.* The determination of a 'deeper' logical representation is described as a later, pragmatic stage of interpretation. For instance,

5 Corresponding, roughly, with what I call *initial logical forms* in this dissertation.

All men are mortal.

[⟨all man⟩ mortal]

($\forall x$: [x human] [x mortal])

Note that in the final stage 'man' has been disambiguated to the sense 'human' rather than 'adult male'. For Episodic Logic, the follow-up to the representation used by Schubert & Pelletier, Schubert & Hwang (1990) presented a more complete description of the generation of sentential logical forms from English.

Knext Schubert (2002) presented the approach of looking beneath the explicit assertional content of text to find knowledge about what relationships and properties are possible in the world. He gives the example that reading 'He entered the house through its open door' suggests that a person – or, at least, a male – may enter a house, a house may have a door, and a door can be open. Knext[6] is an implementation of this extraction based on the compositional semantic analysis of text, *i.e.*, a Montague grammar–inspired semantics along the lines of Schubert & Hwang (1990).

The stages of Knext extraction are:

1 *Preprocess* text, including removing formatting, marking sentence boundaries, and splitting corpora for parallel processing.
2 *Parse* each sentence with a Treebank-trained statistical parser, *e.g.*, that of Collins (1997) or Charniak (2000).
3 *Adjust the phrase structure* for interpretation, *e.g.*, particularizing PP to PP-OF, PP-AT-TIME, *etc.* and replacing SBAR with S-THAT or S-REL. Additional syntactic processing includes correcting systematic phrase attachment errors, assimilating verb particles into the verb, *etc.*

6 The name is a loose acronym for *Kn*owledge *ex*traction from *t*ext and is pronounced either [nɛkst] or [kenɛkst].

4 *Compute logical forms* by applying approximately 80 interpretative rules – regular phrase patterns paired with semantic forms – to compute initial logical forms for the sentence and its constituents in a bottom-up sweep.
5 *Extract and abstract propositions*: Collect phrasal logical forms that may yield stand-alone propositions. The abstraction drops modifiers that are present in the logical forms for lower levels (*e.g.*, adjectival premodifiers of nominals) when constructing LFs for higher levels. Named entities are abstracted as noted below.
6 *Verbalize*: Render the propositions into approximate English.

To abstract to knowledge about classes, Knext uses hand-built *gazetteers* – lists of instances for classes such as *country*, *scientist*, or *male person* – as well as a function to guess the type of a named entity based on patterns such as *Duchess* * ⇒*noblewoman* and * *Co.* ⇒*company*. Factoids produced in this way include *A philosopher may have a conviction*, *A person may say something to a group*, and even epistemic claims like *A person may understand an allure of a part of a book*. The evaluation of Knext extractions is discussed in the following chapter in the context of expanding the selection of textual sources for knowledge extraction.

As a first attempt at producing conditional knowledge for inference from Knext extractions, Van Durme *et al.* (2009) looked at abstracting the possible arguments of a verbal predicate. *E.g.*, given a variety of factoids about things that have wings – eagles, pigeons, planes, hospitals, *etc.* – we want to conclude that *Most things that have wings are birds, planes, or buildings*. These generalizations were made by abstracting up the WordNet nominal hierarchy to find the most specific synsets that cover most arguments. This was done without reference to the frequency with which a factoid is learned, in recognition of the problem of reporting bias (discussed in Chapter 4). However, no logical forms were generated by this initial study, and the feasibility of this abstraction for a large knowledge base is doubtful without further work: Given a full-size Knext KB of tens of millions of factoids, it is prohibitively expensive to find covering hypernyms for

every set of factoids, especially when ambiguous predicates like ‘have’ may take many thousands of distinct arguments.

2.5 Chapter Summary

This chapter has given a selective overview of approaches to knowledge acquisition for artificial intelligence. These include the traditional, manual approach of knowledge engineering, crowdsourced approaches that scale knowledge acquisition from small groups of experts to large numbers of people contributing online, and automated approaches that exploit the availability of text in electronic form. The latter includes work to learn specific relations, arbitrary relations expressed explicitly, and – in the case of Knext – implicit knowledge about what is assumed to be possible in the world. The next chapter investigates the question of what text is suitable for learning general world knowledge with such a tool.

3 Text Sources

Rosencrantz: What are you playing at?

Guildenstern: Words, words. They're all we have to go on.

TOM STOPPARD, *Rosencrantz and Guildenstern are Dead*, 1966

3.1 Introduction

Knowledge extraction efforts have often used corpora of heavily edited writing and sources written to provide the desired knowledge (*e.g.*, newspapers or textbooks). However, the ease of publishing online has created an instantly-available, up-to-date, and increasingly comprehensive store of human knowledge, opinion, and experience. These same features that attract human readership also motivate the use of Web text for automated knowledge extraction.

Traditional corpora will usually possess domain biases that are undesirable for knowledge extraction: Project Gutenberg's collection of public-domain books may contain little knowledge about cellphones but plenty about telegrams; US newswire circa 1999 will have exhaustive knowledge about impeaching a president, but it probably has little that can be learned about dreaming or owning a cat. Rather than construct ever-larger balanced collections of text to use with knowledge-acquisition systems like Knext, we are interested in discovering whether the vast amount of (often) ungrammatically written, unedited, unfocused writing that can be found on the Web can prove an adequate substitute for more formal text resources.

The questions being considered are:

1. Does the volume of extracted general factoids grow indefinitely as more and more weblog sentences are processed (up to hundreds of millions), and similarly as Wikipedia sentences are processed?
2. To what extent do weblog-derived factoids cover Wikipedia-derived factoids and *vice versa*?
3. Does factoid quality depend significantly on the two types of sources?
4. Can extraction quality be significantly improved using a collection of filtering techniques, such as removal of factoids that fail logical-form parsing, violate verb arity constraints, or contain many unlexicalized word stems?

We show that the answers to (1) and (4) are positive and the answers to (2) and (3) are 'less than might be expected'.

3.1.1 *Wikipedia*

Wikipedia is perhaps the most interesting target for current knowledge-extraction efforts, both because of the great diversity of topics it describes and because of its mix of writing styles, ranging from high-profile articles with much-edited language to article stubs consisting of one person's random scribblings, soon to be deleted. As such, it represents a middle ground between the formality of many traditional corpora and the free-for-all nature of weblogs.

Wikipedia articles are written for the express purpose of conveying accurate information about the world, not opinions, anecdotes, *etc.* This might seem to make Wikipedia the obvious best choice for knowledge extraction, but it is a resource for facts stated *explicitly* while Knext targets the general world knowledge that is *implicit* in writing. If weblogs (and similar unstructured, untargeted text, *e.g.*, forum posts) can be of the same utility, they would be a more attractive resource since there is a much greater quantity of such text than Wikipedia articles.

3.1.2 *Weblogs*

As an experiment, we processed the ICWSM 2009 Spinn3r dataset (Burton *et al.*, 2009) of 62 million postings from August–October 2008 to weblogs and other sites that provide RSS/Atom syndication feeds, totalling 203 gigabytes of text. Much of the content included in this dataset is not in English or does not constitute *writing*. Rather, it is the result of people posting pictures, videos, snippets of code, or advertisers' auto-generated text and keywords. The English writing included is rarely straightforward, including song lyrics, sentence fragments littered with emoticons, and unpunctuated train-of-thought. Since the data originates from syndication feeds, many posts are only excerpts and may be truncated mid-sentence.

3.1.3 *Preprocessing*

For these experiments, we used a complete snapshot of English Wikipedia as of 2 July 2009, which was stripped of Wiki markup, links, and figures using a tool by Antonio Fuschetto of the University of Pisa.[1]

To prepare the weblog data for parsing, I stripped the HTML tags marking paragraphs, formatting text, embedding media, *etc.*, eliding text inside of tags whose content is unlikely to be understood without special handling, *e.g.*, ⟨table⟩s, ⟨code⟩ fragments, or ⟨pre⟩formatted text. Many non-English posts are included, for which a statistical parser will blithely generate non-sensical analyses. While much of this was pre-filtered by removing sentences containing frequent words from other languages, it is also dealt with in a post-processing phase that will be described. After this preprocessing, the weblog dataset was reduced to 245,361,917 recognized sentences (26 gigabytes of text) – just 12% of the original data set. This heavy filtering reflects a strong preference for precision over recall.

1 http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

Web text also prompted special handling of URLs and email addresses, substituting 'this website' and 'this email address'. While these replacements oversimplify the ways that these addresses can be used in writing, they allow for sensible extractions from sentences like 'nytimes.com posted an interesting link about…', from which we learn that *A website may post a link*. Additional substitutions correct for common misspellings and accommodate the casual mode of writing often found online, *e.g.*, changing 'u r' to 'you are'. Far from being prescriptivist, this is a necessary step to get a correct syntactic analysis from parsers trained on newswire and other formal writing.

3.2 Rates of Knowledge Extraction

The total number of factoids produced (that is, the number before any filtering) can be seen in Table 3.1 along with the number of factoids produced per 100 words – the *extraction density*. For comparison, the same results are shown for two more traditional corpora: the Brown corpus and the New York Times portion of the Gigaword corpus (Graff *et al.*, 2007). The weblog corpus has a lower extraction density than more formal sources and a higher rate of duplicates, reflecting the noisy nature of much of the writing encountered, including a lack of punctuation and capitalization in many postings, which leads to apparent run-on sentences that are discarded by the parser. On the other hand, the very high rate of unique factoids extracted from the Brown corpus is a reflection both of its topical variation and its very small size.

However, as shown in Figure 3.1, as the number of raw factoids generated increases, the number of unique factoids generated only falls off slightly. This means that there is a fairly consistent benefit to reading more text from each source. Since the amount of weblog text (and other casual, undirected writing on the Web) in existence is vast and continues to grow, a knowledge extraction system like Knext can continue to learn more about the world from the Web almost indefinitely: Any significant fall-off in results won't occur until after many hundreds of millions of sentences are read. While Wikipedia is

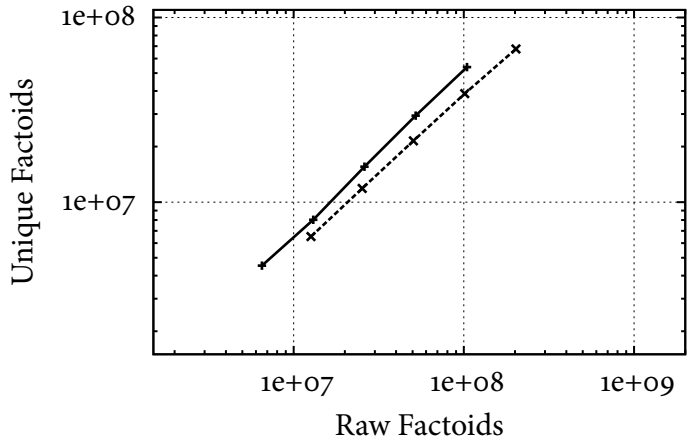

Figure 3.1: **The growth of unique factoids learned from Wikipedia and weblogs as more raw factoids are generated.** The dashed line is for the weblog corpus; the solid line for Wikipedia.

also growing, its standards for worthy topics and for providing sources imply that text is added more slowly; even as new writing is added, other parts are being deleted.

Looking to its anticipated use, we might ask whether a knowledge base that continues to grow indefinitely is a good thing. The answer is a qualified yes: As we continue to acquire more knowledge, the knowledge we haven't seen before is more likely to be about specific individuals or esoteric attributes. Thus there is a declining utility to learning more. However, when we seek to abstract from specific knowledge to more general truths that are unlikely to be stated in text, processing large volumes of text may result in better generalizations.

3.3 Knowledge Overlap

To evaluate the potential usefulness and limitations of extracting from sources like the weblog corpus, we are interested in measuring the types of knowledge that can be learned. For instance, can text that was written without the explicit goal of conveying world knowledge cover what we can learn from Wikipedia? Since we are interested in general world knowledge like *Men have legs* rather than specific facts like *David Bowie was born in 1947*, Wikipedia may not have a decisive advantage.

			Factoids		*Per 100 wds*	
	Sentences	*Words*	*Total*	*Uniq.*	*Total*	*Uniq.*
Weblogs	95,296,872	2,004,492,555	202,282,757	67,632,550	10.1	3.4
Wikipedia	53,971,864	909,756,011	104,287,529	53,945,110	11.5	5.9
NY Times	39,433,116	773,074,059	124,956,881	43,939,886	16.2	5.7
Brown	51,763	1,026,595	132,314	109,443	12.9	10.7

Table 3.1: **The number of factoids extracted from Web and traditional corpora.** Sentence counts are the number of sentences parsed and then used for knowledge extraction, which in the case of the weblogs is smaller than the total available corpus.

Knext generally learns a different set of factoids from weblogs than it does from Wikipedia. Only 5,226,089 unique factoids are found in exactly the same form in the two corpora. This means that just 7% of what we learn from the weblogs can also be found in Wikipedia, and 9.6% of what we learn from Wikipedia can be found in the (larger) set of weblogs.

A sign of how distinct these corpora are: If, after we've extracted from 50 million weblog sentences, we double the corpus to 100 million weblog sentences, that gives a 68% increase in the number of unique factoids. However, if we instead extract from 50 million Wikipedia sentences, we will have a 115% increase in the number of unique factoids. Rather than indicating that Wikipedia is a richer source, this shows that the knowledge it contains generally hasn't been encountered in the weblogs.

However, there are two reasons to doubt that the knowledge found in these corpora is quite this disjoint: (1) There are many differences in diction and spelling that can lead to distinct factoids with nearly identical meanings. For instance, a factoid may be about *a time line* or *a timeline* or a factoid may be about *distances* rather than *a distance*. These representational differences reflect the close connection of the initial logical forms (factoids) to the source material and may be collapsed into a single sharpened logical form

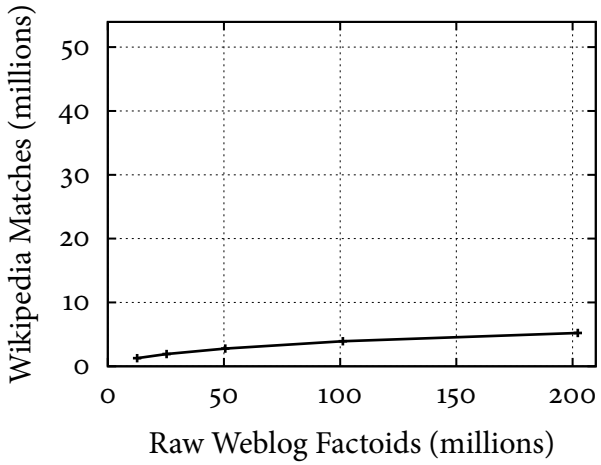

Figure 3.2: **Coverage of Wikipedia factoids by increasing amounts of the raw extractions from weblogs.**

by Lore. (2) Much of the non-overlap data consists of overly specific facts (often about individuals) and factoids seemingly derived from noisy text.

Figure 3.2 shows how many of the Wikipedia factoids can be found in ever larger chunks of the (raw) weblog output. Counting all of the Wikipedia output, we see that the gains made are quite slow, which is unsurprising given that the raw output includes facts about many named entities that could not be abstracted by the current set of gazetteers and are unlikely to receive much discussion on weblogs. There are, *e.g.*, rather few weblog posts about Lucius Seneca.

3.4 Text Sources and Knowledge Quality

In their work evaluating knowledge extracted from the Brown corpus, Schubert & Tong (2003) found substantial differences in judgements of factoid quality depending on the literary style of the source sentences, comparing 'straightforward, realistic narratives in plain, unadorned English' with 'philosophical and theological essays employing much abstract and figurative language', with approximately ¾ of the knowledge from the former being judged reasonable compared with ½ of the latter. How much worse, then, is the

writing we find on the Web, which is eclectic both in topic and writing style? What would it take to produce a collection of good knowledge from such text?

In this section, I introduce a method for filtering the factoids acquired from the Web to find a core set of high-quality knowledge. I offer an assessment of the quality of knowledge that can be learned from unstructured, unedited weblog text and from the more edited, knowledge-oriented writing of Wikipedia – with and without such filtering – and consider whether weblogs could be a worthwhile source for high-quality knowledge mining compared with Wikipedia.

3.4.1 *Automatic Filtering*

The real challenge of Web data is to recognize the subset of useful general world knowledge among the chaff. The factoids we wish to discard include those generated from non-English text remaining in the weblogs, those with multiple uncorrected spelling errors, and those mistakenly generated from all sorts of non-text that failed to be preprocessed away.

To remove factoids generated from remaining non-English text and those generated from sentences with multiple uncorrected spelling errors, we added a post-processing step of checking factoids verbalizations against a lexicon (the contents of WordNet and manual additions), discarding those containing less than 75% known words.[2] Restricting factoids to only using known vocabulary would result in higher quality output but with an unacceptable trade-off in coverage. This cut-off reflects that even non-English sentences may contain words that are also found in English, but we also want to allow for potentially useful propositions containing neologisms that won't be found in our lexicon, *e.g.*, *A blogosphere may explode with discussion*. An example of a proposition that is rejected by this filter is *All mimsy can be borogroves*, learned from a weblog post containing an excerpt from the poem 'Jabberwocky' (Carroll, 1871).

2 In later work, this was changed to allow a maximum of one unknown word per factoid, not counting named entities.

As with less noisy data, errors in the syntactic parsing of English are a common source of bad factoids. For instance, incorrect prepositional phrase attachments in parse trees frequently result in missing arguments, giving incomplete factoids like *A person may feel* where what we want to learn is that *A person may feel an emotion*. To avoid these incomplete factoids, the filter checks whether a predicate's usage matches the range of arities attested in PropBank (Kingsbury & Palmer, 2003) for the corresponding verb. This turns out to be a rather weak restriction given the wide range of possible uses for common verbs, many being uses that Knext is unlikely to output. For instance, PropBank includes a use of *say* with no arguments ('Let's assume someone, say John, has been killed'), while Knext typically encounters it as a transitive verb. A hand-authored set of corrections to these arity ranges limit verbs to their common uses.

Factoids with the vague predicates *thing* or *thing-referred-to* are also removed. This initial study did not include factoids about named entities, and it included a step of checking that output factoids ILFs could be parsed to ensure they were not malformed. For speed, this filtering step was later replaced by simple tree patterns that detect common forms of incorrect logical syntax that can result from some unexpected inputs.

As an estimate of the percentage of each corpus that gets removed by these filtering steps, we ran 2000 randomly selected factoids from each corpus through the filter: 567 (28%) of the weblog factoids and 722 (36%) of the Wikipedia factoids were removed. The greater number of factoids thrown out from Wikipedia stems from the greater number of named entities discussed in Wikipedia that could not be abstracted and were thus removed by the filter as probably being overly specific.

3.5 Evaluation of Knowledge Quality

We are interested not only in *what* we can learn from different Web corpora but also the quality of this knowledge: A large but noisy knowledge base will be of little use in reasoning. To measure the quality of knowledge, we must rely on assessments by human judges.

The statement above is a reasonably clear, entirely plausible, generic claim and seems neither too specific nor too general or vague to be useful:

1 I agree.
2 I lean towards agreement.
3 I'm not sure.
4 I lean towards disagreement.
5 I disagree.

Figure 3.3: Instructions for scaled judging.

We selected 100 propositions uniformly at random from the unfiltered, non-unique[3] output of Knext on each corpus. These were shuffled together and their English-like verbalizations were displayed to the judges – in this case, two of the authors – along with the rating instructions of Van Durme *et al.* (2008), seen in Figure 3.3. Thus the judges did not know which source the factoid they were rating came from nor whether it was among those that would be filtered away.

Some characteristic examples of factoids that were given each rating (agreed on by both judges) are:

1 *A person may have a head.*
  [⟨det person.n⟩ have.v ⟨det head.n⟩]
2 *A thing can be readable.*
  [⟨det thing.n⟩ (be.be readable.a)]
3 *A male may have a call.*
  [⟨det male.n⟩ have.v ⟨det call.n⟩]
4 *Currents can be with some surface electrodes.*
  [(k (plur current.n)) with.p ⟨some-number-of (nn surface.n (plur electrode.n))⟩]

3 Non-unique output was used to favor more frequently generated propositions. No duplicates were selected.

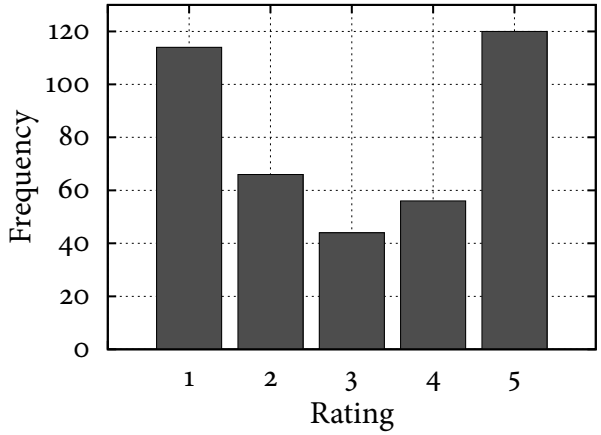

Figure 3.4: **Frequency of ratings assigned to unfiltered factoids from both corpora.** Lower scores are better; see Figure 3.3.

5 *A % may undergo a deflation.*

[⟨det %.n⟩ undergo.v ⟨det deflation.n⟩]

While the highest rated factoid is always true and is at a good level of generality (*person* rather than, say, *male* or *child*), the factoid rated as a 2 is true (some things *are* readable) but is underspecified: What kind of thing is readable? 3 is hard to judge: A person may have a calling or may receive a call, but is the factoid saying either of these? The factoid rated 4 seems a bit too specific (*surface* electrodes) and also a bit vague (*with* them?). The factoid rated 5 we cannot imagine using as knowledge even though we might read a meaning into it: If we take the percent sign to be an adequate stand-in for 'percent', we still don't know what it is a percent of. Factoids at each of these ratings can exhibit different problems, but Van Durme *et al.* (2008) found in the past that judges are less likely to agree *what* it is that's wrong with a factoid than *how good* one is. The distribution of factoid ratings across both corpora can be seen in Figure 3.4.

The assessments Table 3.2 indicate an improvement in the quality of factoids after filtering when compared with the evaluations of the entire, unfiltered set. (For comparison, we estimate that the judgements of Knext's output on the Brown corpus, converted to our current rating scale, would have an average rating of around 2.0. This high rating can largely be ascribed to the accuracy of hand-parses *vs* machine parses.) The evalu-

	Filtered Only			*All*			
	Judge 1	*Judge 2*	*Corr.*	*Judge 1*	*Judge 2*	*Corr.*	*MTurk*
Weblog	2.54	2.52	0.76	3.07	2.98	0.79	2.85
Wikipedia	2.69	2.35	0.71	3.09	2.88	0.76	2.75
Both	2.61	2.43	0.73	3.08	2.93	0.78	2.80

Table 3.2: **Average quality of filtered factoids from Web sources.** Lower scores are better; see Figure 3.3. The last column presents crowdsourced evaluation using Mechanical Turk; see Appendix A.

ations give no indication that the factoids from one Web corpus are of higher general quality than those from the other, with the judges giving roughly the same average rating to each source. A larger sample of 300 factoids from each source was evaluated by non-expert judges on Amazon Mechanical Turk. They rated Wikipedia factoids a bit better, and overall assessed quality as higher than the expert judges. For details on this evaluation method, see Appendix A.

Beyond this filtering, we can also consider only including factoids that are found more than once. Van Durme *et al*. (2008) found that propositions that were extracted at least twice were, on average, judged to better than those extracted only once. However, as extraction frequency continued to increase, the level of judged acceptability did not. We found that for the 200 factoids that were rated, those extracted only once were rated 3.4 on average, while those rated twice or more often were rated 2.79 on average. This is slightly less effective than the other filtering techniques alone. Combining the two, we get a filtered subset of factoids with an average rating of 2.34 *vs* 3.00 overall.

3.6 Chapter Summary

When extracting general world knowledge, does it matter what machines read? Our findings are that:

1 The intuition that casually written material on the Web may be less useful for general knowledge mining than more formal sources like Wikipedia is to some extent confirmed by the lower extraction rates we achieved using Knext on weblogs.
2 For both sources, the volume of unique extracted general factoids grows indefinitely, with little sign of leveling off on a logarithmic scale, even after processing of hundreds of millions of weblog sentences.
3 Wikipedia-derived general factoids cover only a small fraction of weblog-derived facts and the converse holds also, though the coverage of Wikipedia-derived factoids by weblog-derived factoids appears to grow indefinitely.
4 Despite the different writing quality in weblogs and Wikipedia, the quality of extracted propositions from those sources are rated about the same by human judges
5 The use of multiple filtering techniques, such as removal of propositions that fail logical-form parsing, or violate verb arity constraints, or contain many unlexicalized word stems, significantly improves the quality of extracted propositions.

Our results suggest that general knowledge extraction from Web-scale text, supplemented with automatic filtering, has the potential to produce large, symbolic knowledge bases of good quality, as judged by people. The next chapter questions standard assumptions about the use of text and textual frequencies to acquire knowledge that is representative of the world.

4 Reporting Bias

> The aspects of things that are most important for us are hidden because of their simplicity and familiarity.
>
> LUDWIG WITTGENSTEIN, *Philosophical Investigations*, 1953

Much work in knowledge extraction from text tacitly assumes that the frequency with which people write about actions, outcomes, or properties is a reflection of real-world frequencies or the degree to which a property is characteristic of a class of individuals. In this chapter, I question this idea, examining the phenomenon of *reporting bias* and the challenge it poses for knowledge extraction. I conclude with discussion of approaches to learning commonsense knowledge from text despite this distortion.

4.1 Introduction

A system can look for explicit assertions of general knowledge or knowledge implicit in recurrent patterns of predication and modification, or it can abstract general claims from collections of specific instances. Regardless of the *modus operandi*, it is necessary to distinguish knowledge about what *normally* holds in the world from the atypical or claims that are simply not true. For instance, the Knext (Schubert, 2002) system for knowledge extraction from text (described in § 4.4) learns both *The Earth may revolve around the Sun* and *The Sun may revolve around the Earth*. Mistaken claims like the latter may indicate a failure to correctly learn from text (*e.g.*, if a source said 'It is *not* the case that the

Earth revolves around the Sun'), or it may result from reading an inaccurate or fantastical text.

To identify good claims, it is typical to take an inductive view, with textual references serving as evidence: The more often we read something, the more likely it is to reflect what is true in the real world. This is intuitively reasonable, and, over a large collection of texts, Knext learns the heliocentric claim 327 times, while the geocentric claim is only learned 126 times. However, the frequency with which situations of a certain type are described in text does not always correspond to their relative likelihood in the world, or even the subjective frequency captured in human beliefs. For instance, from the same texts, Knext learns almost a million times that *A person may have eyes*, but fewer than 1,600 times that *A person may have a spleen*. While eyes are discussed frequently, many other body parts are not – but this doesn't mean they're any less common in people. We will refer to this potential discrepancy between reality and its description in text as *reporting bias*.

For knowledge extraction (KE), we are interested in reporting bias as it relates to the frequency with which events or actions occur, the likelihood of specific outcomes, and the prevalence of properties. If our textual examples are not representative of reality, then claims induced from them are likely to be inaccurate. For instance, according to Douglas Lenat, at one point Cyc 'concluded that everyone born before 1900 was famous, because all the people that it knew about and who lived in earlier times were famous people.' (Moody, 1999)

While the focus of this discussion is on how reporting bias affects the acquisition of general knowledge, many of the phenomena we discuss also apply to factual information extraction (IE). *E.g.*, frequently reading claims that Barack Obama was born in Kenya does not make it a reliable extraction. However, for a factual IE system, other extraction properties may be more salient than textual frequency. For instance, the great frequency of statements that George Bush is the president of the United States should not lead us to believe this is currently true, given the greater recency of sentences indicating Obama

is president. The trustworthiness of text sources can also be of greater importance for factual IE than for general knowledge extraction, which can abstract claims even from realistic fiction.

§ 4.2 presents evidence of reporting bias by contrasting frequencies found in text and in the world. §4.3 proposes an explanation of reporting bias as a systematic distortion of reality. § 4.4 looks at how reporting bias affects some existing knowledge-extraction systems and at attempts to correct for it. § 4.5 suggests approaches for future work.

4.2 Measuring Reporting Bias

To demonstrate the reality of reporting bias and motivate our discussion in the next section, we will give several examples where the frequencies of textual references and extractions differ significantly from what we know to be the case in the world. Giving a full, accurate model of reporting bias or establishing how widespread the problem is would require the availability of real-world frequencies across the range of types of properties that we are interested in learning from text. Instead, we simply demonstrate the existence of significant reporting bias for actions or events, outcomes, and properties.

We present textual frequencies based on the Google Web 1T n-gram data (Brants & Franz, 2006), which is derived from approximately a trillion words of Web text. We support this, where possible, with the number of times Knext learns a relevant claim about the world. These results are taken from a knowledge base of 73 million unique factoids learned from sources including the Brown Corpus (Kučera & Francis, 1967), the British National Corpus (BNC Consortium, 2001), Gigaword (Napoles *et al.*, 2012), electronic books from Project Gutenberg, Wikipedia, and the ICWSM 2009 weblog corpus (Burton *et al.*, 2009).

In the introduction, we used the example of how often we are told a person has spleen *vs* having eyes. In Table 4.1, we see the significant variation with which body parts are mentioned in writing, though they are near universally present in individuals. While this type of knowledge is readily available from manually created sources such as Word-

Body Part	*Teraword*	*Knext*	*Body Part*	*Teraword*	*Knext*
Head	18,907,427	1,004,300	Liver	246,937	9,452
Eye(s)	18,455,030	934,721	Kidney(s)	183,973	3,289
Arm(s)	6,345,039	399,120	Spleen	47,216	1,568
Ear(s)	3,543,711	309,708	Pancreas	24,230	1,186
Brain	3,277,326	144,511	Gallbladder	17,419	991

Table 4.1: **Textual support for body-part extractions.** N-gram frequencies are for (*his*|*her*|*my*|*your*) ⟨*body part*⟩ and the number of times Knext learns *A* ⟨*body part*⟩ *may pertain to a person*. Plurals are included when appropriate.

City	*Population*	*Google*	*Ref./Pop.*
New Delhi	21,750,000	8,130,000	0.37
Beijing	20,180,000	1,630,000	0.08
NYC	8,245,000	16,000,000	1.94
London	8,174,000	33,700,000	4.12
Toronto	2,615,000	8,490,000	3.25
Detroit	706,585	5,880,000	8.32

Table 4.2: **Textual support for extractions about city populations.** Google results are for 'lives in ⟨city⟩'. For 'NYC', we count 'New York City' or 'NYC', but not more specific terms. Population figures are Wikipedia's report of appropriate census results from 2011.

Net (Fellbaum, 1998) or Cyc (Lenat, 1995), the fact that even such simple extractions exhibit significant reporting bias bodes ill for the long tail of more subtle knowledge that we are less likely to be able to enumerate.

For instance, KE systems may try to learn from text the typical frequency of an event or how characteristic an action is of a class of individuals, to produce generic claims such as *Generally people sleep* or *Most people sleep daily*, while only *Some people play the viola*. However, in Table 4.3, we see that murder is mentioned in text many more times than more quotidian actions like hugging or constant activities like breathing, and we find people are late much more than they are on time. The Knext extraction frequencies can be seen as a further distortion of the textual frequencies, due, at least in part, to the filter-

Word	*Teraword*	*Knext*	*Word*	*Teraword*	*Knext*
spoke	11,577,917	372,042	hugged	610,040	11,453
laughed	3,904,519	179,395	blinked	390,692	21,973
murdered	2,843,529	16,890	was late	368,922	31,168
inhaled	984,613	5,617	exhaled	168,985	4,052
breathed	725,034	41,215	was on time	23,997	14

Table 4.3: **Textual support for extractions about verbal events.** N-gram frequencies are for the verb alone. Knext counts are the number of times it learns that *A person may* ⟨*x*⟩, including appropriate arguments, *e.g.*, *A person may hug a person*.

ing of potential claims. For instance, factoids about murder are automatically discarded if they lack the complement (*i.e.*, you need to murder *someone*). Additionally, the BLLP parser systematically misparsed 'murdered' as a passive (VBN) rather than simple past tense, even in simple sentences like 'Brutus murdered Cæsar'.

Another important kind of knowledge is the expected outcome of an action or event, *e.g.*, *If a person drops a glass, it may break*. As this knowledge relies on larger patterns of predication, often involving more than one sentence, it is not easily measured on a large scale. However, in Table 4.5 we see that, per mile travelled, a person is more likely to experience a crash on a motorcycle than in a car or in an airplane. However, in text motorcycle crashes are only mentioned half as frequently as plane crashes.

For a simpler example, we know that for most races (whether foot races, political contests, *etc.*) the number of winners is less than or equal to the number of losers, yet we find far more reports of a person winning a race than losing it: In the n-grams, *won the race* occurs more than six times as often as *lost the race* (66,011 *vs* 10,430). The number of matches for (*participated in*|*ran in*|*took part in*|*entered*) *the race*, which lack the stigma of 'losing', is still quite low (22,512). Even for the Academy Awards, where 'it's an honor just to be nominated', people are much more likely to write about a win than a nomination. We find *won the academy award* 15,098 times *vs* 4,551 for *nominated for the academy*

Phrase	*Teraword*
won the academy award	15,098
nominated for the academy award	4,551
academy award winning	45,953
academy award nominated	10,833
academy award winner	66,244
academy award nominee	22,512

Table 4.4: **Textual support for winning an Academy Award *vs* being nominated.** While more actors and films are nominated for academy awards than win them, text is more likely to mention wins.

Type	*Miles Travelled*	*Crashes*	*Miles/Crash*	*Teraword*
Car	1,682,671 million	4,341,688	387,562	1,748,832
Motorcycle	12,401 million	101,474	122,209	269,158
Airplane	6,619 million	83	79,746,988	603,933

Table 4.5: **Textual references to vehicular accidents.** *Miles Travelled*, *Crashes*, and *Miles/Crash* are for travel in the United States in 2006 (US Department of Transportation, 2009). A plane crash is considered any event in which the plane was damaged. *Teraword* results are for the patterns *car* (*crash*|*accident*), *motorcycle* (*crash*|*accident*), and (*airplane*|*plane*) (*crash*|*accident*).

award (and the same is true for a number of variations, such as *academy award winner* and *academy award winning*) – see Table 4.4.

In Table 4.2, we show that the number of times we read that a person lives in a city is quite disproportionate to the number of people who actually live in the city. This effect is strongest for cities in non-English-speaking countries, where we see fewer references than residents. (The lower reporting for New York City than for London, Toronto, and Detroit may be due to the tendency to indicate a specific borough ('Brooklyn') or neighborhood ('SoHo') when talking about NYC or to refer to it simply as 'New York', which is ambiguous without context.)

4.3 Discussion

We believe these discrepancies between reality and textual frequency indicate a pervasive distortion. Reporting bias results from our responsibility as communicators to be maximally informative in what we convey to other people, who share our general world knowledge, and to convey information in which they are likely to be interested.

The first of these imperatives was postulated by Paul Grice (Grice, 1975) as his *conversational maxim of quantity*. This states that communication should be as informative as necessary – but no more, leaving unstated information that can be expected to be known or can be inferred from what is said using commonsense knowledge. Clark (1975) observed that while Gricean implicatures are linguistic – are part of an intended message – they draw not just on linguistic knowledge but on knowledge of objects and events in the world. Havasi *et al*. (2007) previously connected knowledge acquisition from text to Gricean principles, noting that 'people tend not to provide information which is obvious or extraneous' and, therefore, 'it is difficult to automatically extract common-sense statements from text, and the results tend to be unreliable'. The second imperative – to be interesting – is less a linguistic principle than a psychological or social one: Some topics are intrinsically interesting to people, regardless of their prevalence, and we will tend to discuss these, biasing what information is available in text.

To elaborate and clarify this discussion, we offer these hypotheses about reporting bias with corresponding examples:

1 *The more expected something, the less likely people are to convey it as the primary intent of an utterance.*

People are unlikely to tell you about 'the man with two legs' or 'a yellow pencil'. Rather, we state exceptional properties: 'a man with one leg', 'a blue pencil'. Similarly, we don't say 'I paid for the book and then I owned it' or 'A suicide bomber blew himself up yesterday. He died.' as these are assured consequences. We might, however, say, 'I crashed my car. It was totalled.' as the degree of damage is not certain otherwise. While expected information

is unlikely to be the *primary purpose* of an utterance, it can appear in presuppositions; see §4.5.

2 *The more value people attach to something, the more likely they are to give information about it, even if the information is unsurprising.*
For instance, in a report of forest fires sweeping parts of California, we care about homes destroyed, and people killed or injured, but most care less about the number of chipmunks or deer killed. Further, the destruction of thousands of acres of forest will often matter and will be mentioned, as would the loss of members of a rare animal species. If we describe a person we met, we may well say he has brown hair even though this extremely common. However, we are even more likely to mention a person's hair color if it's unusual: While textual references to brown hair are more frequent than red (594,997 to 382,989 in the Google n-grams), the latter's representation is quite disproportionate to its occurrence in the population.

3 *Conversely, even unusual facts are unlikely to be mentioned if they are trivial.*
E.g., having a scratch on the left bicep may be less common than an interesting, important property like a woman being pregnant, but it usually matters too little to be reported.

4 *Reporting bias varies by literary genre.*
There will be considerable differences in the frequency of reporting events in an encyclopedia *vs* in fiction or even, *e.g.*, among different newspapers. While sports pages will 'over-report' sporting events compared to crimes, celebrity shenanigans, or business news, the National Inquirer or the Wall Street Journal might over-report other types of events.

5 *There are fundamental kinds of lexical and world knowledge that are needed for understanding and inference that don't get stated in text.*
This can be because they are innate or are learned before language is acquired, by physical interaction in the world. *E.g.*, physical objects can't be in different places at the same

time; solid objects tend to persist (in shape, color and other properties) over time; if *A* causes *B* and *B* causes *C* then it's usually fair to say that *A* causes *C*; people do and say things for reasons – to get food or possessions or pleasure, to avoid suffering or loss, to provide or solicit information, *etc.*; you can't grab something that's out of reach; you can see things in daytime that are nearby and are not occluded; people can't fly like birds or walk up or through walls; *etc.*

There are also the lexical entailments and presuppositions that we learn as part of language and hardly ever say: 'above' and 'below', 'bigger' and 'smaller', 'contained in' and 'contains', 'good' and 'bad', *etc.*, are incompatible; dying entails becoming dead; going somewhere entails a change in location; walking entails moving one's legs, *etc.*

4.4 Previous Approaches

In looking at how systems have dealt (or not dealt) with reporting bias, we want to contrast three lines of work: information extraction systems (Cowie & Lehnert, 1996; Sarawagi, 2008), which learn explicitly stated material; knowledge extraction systems (*e.g.*, Schubert, 2002), which abstract individual instances to the general knowledge that's implicit in them; and systems that learn general rules implicit in a collection of specific extractions (*e.g.*, Raghavan & Mooney, 2013; Van Durme *et al.*, 2009; Carlson *et al.*, 2010). We only provide a few examples; for a more thorough overview, see Chapter 2.

TextRunner TextRunner (Banko *et al.*, 2007) is a tool for extracting explicitly stated information as tuples of normalized text fragments. Its output includes both information about specific individuals and generic claims. Based on the number of distinct sentences from which a tuple was extracted, it is assigned a probability of being a correct instance of the relation. TextRunner's authors view the probabilities assigned to these claims not as representing the real-world frequency of an action or the likelihood the relation holds for an instance of a generic subject, but simply as the probability that the tuple is 'a correct instance of the relation'. It's not clear what this means for their 'abstract tuples', which are

86% of the output on average, per relation, and include claims such as *(Einstein, derived, theory)* or *(executive, hired by, company)*. Is this a correct instance if Einstein at any point derived a theory? What if any executive was at some point hired by a company? Or is an abstract tuple only a correct instance of the relation if it is a good generic statement, *e.g.*, *Executives are (generally) hired by companies*?

Knext Knext (Schubert, 2002; Van Durme & Schubert, 2008), under development since before 2002, is a tool for extracting general world knowledge from large collections of text by syntactically parsing each sentence with a Treebank-trained parser (*e.g.*, Charniak, 2000) and compositionally applying interpretive rules to compute logical forms in a bottom-up sweep, abstracting those that serve as stand-alone propositions.The results are quantificationally underspecified Episodic Logic formulas, which are verbalized in English as possibilistic claims, *e.g.*, *Persons may want to be rid of a dictator*. Knext treats all discovered formulas as *possible* general world knowledge. In an evaluation of 480 propositions, Van Durme & Schubert (2008) observed that propositions found at least twice were judged more acceptable than those extracted only once. However, as the support increased above this point, the average assessment stayed roughly the same. That is, frequency of extraction was not found to be a reliable indication of quality.

Urns TextRunner's probabilities use the Urns model of Downey *et al.* (2005, 2010), which is based on the belief that an extraction is more likely to be true if it is obtained from multiple documents, adjusting for how often a type of reference occurs. *E.g.*, Urns should assign a lower probability to 'countries such as Washington' (31 hits on the Web) than it does to 'throwable objects such as bean bags' (3 hits) given the far greater number of extractions for countries than for throwable objects (example due to Doug Downey). However, Urns is meant to establish the truth of ground facts (*e.g.*, *Einstein was born in 1879*), not the probability of a generic claim applying (*e.g.*, *People eat food*). Indeed, a great deal of the commonsense knowledge we want to learn is only discovered a handful

of times, even over Web-scale text, while Urns requires a fairly large sample size for each relation.

Learning Rules from Extracted Facts A line of work at Oregon State University (Sorower *et al.*, 2011; Doppa *et al.*, 2010) learns domain-particular rules based on specific facts extracted from text. They address a subproblem of the general reporting-bias phenomenon, namely the conditional bias of our Hypothesis 1. If attribute $A(x) = a$ of some entity is reported, and $A(x) = a$ tends to imply $B(x) = b$, then $B(x) = b$ tends not to be reported. (*E.g.*, if someone is stated to be a Canadian citizen, then we are less likely to also state that they were born in Canada.) But if, in fact, $B(x) = b'$, then we are likely to say so. (*E.g.*, we *would* say 'an Egyptian-born Canadian'.)

Along similar lines, Raghavan & Mooney (2013) learn commonsense knowledge in the form of probabilistic first-order rules from the incomplete, noisy output of an information-extraction system. Their rules have a body containing relations that are often stated explicitly, while the head uses a relation that is mentioned less often as it's easily inferred. They produce rules like $\text{hasBirthPlace}(x, y) \wedge \text{person}(x) \wedge \text{nationState}(y) \Rightarrow \text{hasCitizenship}(x, y)$. An interesting aspect of their approach is the use of WordNet similarity to weight rules, based on the idea that more accurate rules usually have predicates that are closely related in meaning.

4.5 Addressing Reporting Bias

We've shown that reporting bias's distortion of real-world frequency in text makes it doubtful that we can interpret the number of textual references or explicit statements supporting a general claim as directly conveying real-world prevalence or reliability. While there seems to be no silver bullet, there are some approaches to learn what normally holds in the world, several of which are explored in more detail in Chapter 5. For instance, we can focus extraction on more informative constructions:

1 *Presuppositions*. Commonsense knowledge that is rarely stated explicitly can nonetheless appear in sentences as presuppositions – beliefs the speaker expects others to share:

> Both my legs hurt.
>
> ⇒ *A person normally has two legs.*
>
> I forgot the money to buy groceries.
>
> ⇒ *A person may use money to buy things.*

2 *Disconfirmed expectations*. Gordon & Schubert (2011) learned commonsense inference rules from constructions that indicate a speaker's expectation about the world was not met, *e.g.*,

> Sally crashed her car into a tree but wasn't hurt.
>
> ⇒ *If a person crashes her car, she may be hurt.*
>
> I dropped the glass, but it didn't break.
>
> ⇒ *If a person drops a glass, it often will break.*

Other sentences suggest that an action or event has not taken place with the normal temporal frequency (Gordon & Schubert, 2012):

> I hadn't slept in days.
>
> ⇒ *A person normally sleeps at least daily.*

(These claims are implicitly conditioned on whether the agent does the action at all, *e.g.*, *If a person writes a book at all, he probably does so every few years.*)

3 *Implicit denials.* Explicit statements, pragmatically required to be informative, contain implicit denials that what they're saying is usually the case, *e.g.*,

> The tree had no branches.
>
> ⇒ *Trees usually have branches.*

However, these vary in how easily they can be transformed into general claims, *e.g.*,

Molly handed me a blue pencil.

⇒ *Probably pencils are not always blue.*

4 *Reference to individuals*. Expected properties can be expressed when identifying a particular individual, *e.g.*,

…the man I met yesterday.

⇒ *A person may meet a man.*

Claims frequently learned from such constructions may be more usual than those learned from more explicit assertions, though there are still many more references to a 'plane that crashed' than a 'plane that landed'.

More correlation might be seen between frequency and extraction quality if we only count the frequency of distinct textual references. *E.g.*, repeated mentions of the film *True Lies*, misparsed as a common noun phrase, lead Knext to learn *Lies may be true*. Even if text is analyzed correctly for its surface meaning, it can lead to bad knowledge, *e.g.*, the idiom 'when pigs fly' gives us *Pigs may fly*. A related problem is frequently repeated text, such as song lyrics on the Web. To account for this textual bias – exact repetition – we might give more weight to knowledge learned from different extraction methods or just from distinct constructions.

Another possibility is to use a hybrid approach to knowledge extraction, along the lines of Snow *et al.* (2008) or Hoffman *et al.* (2009). For instance, we might combine text mining with a crowdsourced rating (Appendix A) or filtering stage to assign an approximate real-world frequency to the knowledge found most frequently in text. Work in the emerging 'grounded language' movement may also be important. If one were to say 'John entered the room', they are unlikely to follow it up with 'He blinked. He breathed.' However, many mundane actions and activities might be recognized, *e.g.*, by sampling video and be incorporated into our knowledge.

It is also important to recognize that for some problems, frequencies for the distorted world described in text are more useful than real-world frequencies. For instance, a

parser is concerned with how frequently 'cat' is the subject of 'meow', rather than how frequently cats actually meow. With the bias for the interesting or unusual, textual frequencies may also be useful for guiding inference for conclusions that are most likely to be important or useful: If we are told 'John is a person', we don't want to reason that he has skin cells (although this is certainly true) but rather that he probably has a job of some kind, that he lives somewhere, *etc.*

4.6 Chapter Summary

We have argued that researchers need to be aware that the frequency of occurrence of particular types of events or relations in text can represent significant distortions of real-world frequencies and that much of our general knowledge is never alluded to in natural discourse. We provided a brief pragmatic argument for why reporting bias exists, which led to suggestions on how we might, partially, work around it, which are explored further in the next chapter.

Our examples and discussion are meant to provoke further study. If reporting bias is not a real problem for knowledge acquisition, it remains for the community to show this to be the case. Otherwise, more work is called for to determine if, and how, we can correct for it. At worst, reporting bias may prove an upper bound on the extent to which human knowledge can be learned from text and may provoke further work on hybrid approaches to knowledge acquisition.

5 Lore: Learning & Sharpening Implicit Knowledge

> How much do we know at any time? Much more, or so I believe, than we know we know!
>
> AGATHA CHRISTIE, *The Moving Finger*, 1942

This chapter presents work to acquire commonsense knowledge from text. We abstract from particular references to general possibilities and from normative constructions such as disconfirmed expectations to the underlying presumptions and expectations. The resulting knowledge undergoes a process of sharpening by lexical-semantic patterns to produce appropriately strong, partially disambiguated probabilistic inference rules.

This knowledge-extraction system is Lore. A shamelessly selective excerpt from the Oxford English Dictionary (Simpson, 2013) suggests the appropriateness of the name:[1]

> **lore**, *n.*
>
> 1 The act of teaching; the condition of being taught; instruction, tuition, education. In particularized use: A piece of teaching or instruction; a lesson…
>
> 4 …Something that is spoken; information; story; language.
>
> 5 *a.* That which is learned …Also, in recent use, …the body of traditional facts, anecdotes, or beliefs relating to some particular subject…
>
> *b.* A body of knowledge, a science.

1 'Lore' is also familiar to fans of *Star Trek: The Next Generation* as the evil twin brother of the android Data.

In the following sections, I describe in more detail the linguistic constructions from which Lore extracts knowledge and the ways in which it refines these extractions, but first give an overview of the entire knowledge-extraction process:

Phase 1: Syntactic Analysis A text is syntactically parsed using a standard statistical parser.[2] The Treebank-style parse trees are processed with tree transduction rules written in TTT (Purtee & Schubert, 2012) to repair common parse errors and to simplify later interpretation.

Phase 2: Form Factoids from Text Semantically underspecified initial logical forms (ILFs) are formed by two applications of interpretation and abstraction rules: First, a bottom-up application of compositional rules builds potential predications for each constituent up to the sentence level. (This is an augmented version of the traditional Knext (Schubert, 2002) extraction.) Second, application of tree-wide interpretation rules match patterns of predication that are lexicalized or cross constituent boundaries (described in §5.1). Each sentence may generate several ILFs, which express self-contained pieces of knowledge implicit in the sentence. ILFs are filtered to remove those that are missing required arguments or are otherwise malformed (along the lines described and evaluated in §3.4.1).

Phase 3: Form Axioms from Factoids Low-frequency factoids (ILFs) are discarded, with the frequency cut-off weighted by the subjective trustworthiness of the source text. Based on semantic categories (*e.g.*, *agent*, *movable object*, *location*), lexico-semantic properties (*e.g.*, individual- *vs* stage-level predicates), and extraction frequencies, the remaining factoids are transformed into explicitly quantified, partially disambiguated, probabilistic axioms (§5.2).

2 At present this is the Charniak (2000) BLLIP parser with the self-trained model of McClosky *et al.* (2006), but there is no strong commitment to this choice.

Beneath the assertional content of a sentence are presumptions about the kinds of properties, relationships and events that commonly occur in the world. For instance, describing someone as 'ready to bolt like a frightened rabbit at the first sign of condemnation'[3] includes the presumption that rabbits can be frightened. (A reader would balk at comparing someone to a 'frightened wall' since we don't share the presumption that a wall can be frightened.) Generally speaking, a presumption is anything where, in conversation, the listener could object, 'Wait a minute – I didn't know that ϕ!'[4] Schubert (2002) introduced an approach to abstracting presumptive knowledge from text, which was implemented as Knext, the source of extractions in previous chapters. Knext applies rules for compositional semantic interpretation and abstraction, forming general factoids from that syntactic tree. For instance,

[⟨det rabbit.n⟩ frightened.a]

where the angle brackets denote unscoped quantification. Knext automatically verbalizes these factoids into English, expressing the weak, possibilistic meaning of the statement, *e.g.*, 'A rabbit can be frightened'. We first consider augmentations of this extraction method to learn more possibilistic commonsense knowledge from text and then turn to the problem of making this usable for inference.

Named Entities & Abstraction To abstract named entities to classes of individuals, Knext used 55 gazetteers (as well as lists of common given names). In Lore this was replaced with the set of unambiguous class instances (Miller & Hristea, 2006) in WordNet – *i.e.*, all lemmas that only exist as an *instance-of* one or more classes. If a name is an instance of multiple classes, all possible abstractions are generated, so a sentence about Sartre will produce potential claims about a *dramatist#1* and an *existentialist#1*. If a word is ambiguous between multiple senses but they are all instances of the same synset, it

3 Example from the ICWSM 2009 Spinn3r weblog corpus (Burton *et al.*, 2009).

4 A test employed by von Fintel (2004) – adapted from Shannon (1976) – for a broadened notion of *presupposition*, applied to knowledge extraction by Van Durme (2010).

is also included in Lore's abstractions. *E.g.*, 'battle of Ypres' is ambiguous among three distinct battles of World War I, but in each case the correct abstraction is to *pitched battle#1*. In future, these abstractions could easily be expanded further based on Freebase or any number of automated instance-extraction systems. However, the current set of over 1,000 abstraction classes is subjectively judged sufficient for broad coverage.

In addition to abstracting from claims about named entities to their known classes, Lore preserves the specific claim. While most commonsense knowledge is about classes, there is good general knowledge that is about individuals, *e.g.*, *Saturn has rings* or *Delhi is crowded*.

Learning Presumed Numbers Commonsense knowledge that is rarely stated explicitly can nonetheless appear in sentences as presuppositions – beliefs the speaker expects others to share. For instance, while from the sentence 'Both my legs hurt' or the phrase 'my other leg', Lore learns that *A person may have two legs* (which, based on repetition of such forms, is strengthened to the claim that *All or most people have two legs as body parts*). From a construction like 'all his friends', Lore learns that *Three or more friends may pertain to a male*.

5.1 Conditional Knowledge from Text

Reasoning about ordinary human situations and activities requires the availability of diverse types of knowledge, including expectations about the probable results of actions and the lexical entailments for many predicates. This section describes work to acquire such a collection of conditional (if–then) knowledge by exploiting presumptive discourse patterns (such as ones involving 'but', 'yet', and 'hoping to') and abstracting the matched material into general rules.

5.1.1 *Introduction*

We are interested, ultimately, in enabling an inference system to reason forward from facts as well as backward from goals, using lexical knowledge together with world knowledge. Creating appropriate collections of general world knowledge to support reasoning has long been a goal of researchers in Artificial Intelligence. Efforts in information extraction, *e.g.*, Banko *et al.* (2007), have focused on learning base facts about specific entities (such as that Barack Obama is president), and work in knowledge extraction, *e.g.*, Knext, has found generalizations (such as that a president may make a speech). However, even when the meaning of such claims is sharpened to support inference (as in §5.2), these resources don't provide a basis for saying what we might expect to happen if, for instance, someone crashes their car.

That the driver in a car crash might be injured and the car damaged is a matter of common sense, and, as such, is rarely stated directly. However, it can be found in sentences where this expectation is disconfirmed: 'Sally crashed her car into a tree, but she wasn't hurt.' We have been exploring the use of lexico-syntactic discourse patterns indicating disconfirmed expectations, as well as people's goals ('Joe apologized repeatedly, hoping to be forgiven'). The resulting rules are a step toward obtaining classes of general conditional knowledge typically not obtained by other methods.

5.1.2 *Method*

In an initial study, Gordon & Schubert (2011), parse trees were matched using hand-authored lexico-syntactic rules for TGrep2 (Rohde, 2001), centered around pragmatically significant cue words such as 'hoping to' or 'but didn't'. In the current version of Lore, these patterns have been converted to TTT (Purtee & Schubert, 2012) rules that output initial logical forms, with constituents such as noun and verb phrases being passed to Knext's compositional interpretation and abstraction methods. This section describes the initial study of these patterns and their assessment.

Disconfirmed Expectations These are sentences where 'but' or 'yet' is used to indicate that the expected inference people would make does not hold. In such cases, we want to flip the polarity of the conclusion (adding or removing 'not' from the output) so that the expectation is confirmed. For instance, from

> The ship weighed anchor and ran out her big guns, but did not fire a shot.

we get that the normal case is the opposite:

> *If a ship weighs anchor and runs out her big guns, then it may fire a shot.*

Or for two adjectives, 'She was poor but proud':

> *If a female is poor, then she may not be proud.*

Contrasting Good and Bad A different use of 'but' and 'yet' is to contrast something considered good with something considered bad, as in 'He is very clever but eccentric':

> *If a male is very clever, then he may be eccentric.*

If we were to treat this as a case of disconfirmed expectation as above, we would have claimed that 'If a male is very clever, then he may not be eccentric'. To identify this special use of 'but', we consult a lexicon of sentiment annotations, SentiWordNet (Baccianella *et al.*, 2010). Finding that 'clever' is positive while 'eccentric' is negative, we retain the surface polarity in this case.

For sentences with full sentential complements for 'but', recognizing good and bad items is quite difficult, more often depending on pragmatic information. For instance, in

> Central government knew this would happen but did not want to admit to it in its plans.

knowing something is generally good while being unwilling to admit something is bad. At present, we don't deal with these cases.

Expected Outcomes Other sentences give us a participant's intent, and we just want to abstract sufficiently to form a general rule:

> He stood before her in the doorway, evidently expecting to be invited in.
>
> *If a male stands before a female in the doorway, then he may expect to be invited in.*
>
> Elisabeth smiled, hoping to lighten the conversational tone and distract the Colonel from his purpose.
>
> *If a female smiles, then she may hope to lighten the conversational tone.*

While most general rules about 'a male' or 'a female' could instead be about 'a person', there are ones that can't, such as those about giving birth. The raising of terms is left for later work, discussed in §7.3.

5.1.3 *Evaluation*

Initial development of these rules was based on examples from the (hand-parsed) Brown Corpus and the (machine-parsed) British National Corpus. These corpora were chosen for their broad coverage of everyday situations and edited writing. As the examples in the preceding subsections indicate, rules extracted by our method often describe complex consequences or reasons, and subtle relations among adjectival attributes, that appear to be quite different from the kinds of rules targeted in previous work (*e.g.*, that discussed by Sekine, 2008).

For evaluation, we used a corpus of personal stories from weblogs (Gordon & Swanson, 2009). We sampled 100 output rules and rated them on a scale of 1–5 (1 being best) based on the criteria in Figure 3.3. To decide if a rule meets the criteria, it is helpful to imagine a dialogue with a computer agent. Told an instantiated form of the antecedent, the agent asks for confirmation of a potential conclusion. *E.g.*, for

> *If attacks are brief, then they may not be intense*,

Judge 1	*Judge 2*	*Correlation*
1.84	2.45	0.55

Table 5.1: **Average ratings and Pearson correlation for rules.** Lower ratings are better; see Figure 3.3.

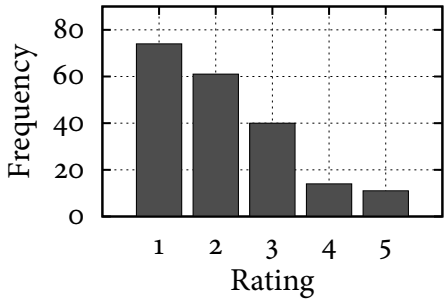

Figure 5.1: **Counts for how many rules were assigned each rating by judges.** Lower ratings are better; see Figure 3.3.

the dialogue would go:

— The attacks (on Baghdad) were brief.

— So I suppose they weren't intense, were they?

If this is a reasonable follow-up, then the rule is probably good, although we disprefer especially unlikely antecedents – rules that are vacuously true.

As the results in Table 5.1 and Figure 5.1 indicate, the overall quality of the rules learned is good but there is room for improvement. We also see a rather low correlation between the ratings of the two judges, indicating the difficulty of evaluating the quality of the rules, especially since, for the initial experiment, their expression in natural language (NL) makes it tempting to 'fill in the blanks' of what we understand them to mean. The difficulties of judging inferential rules in isolation – even when they are expressed in English – motivates the evaluation of simple inferences made with them in Chapter 6.

Rules that both judges rated favorably (1) include:

If a pain is great, it may not be manageable.

If a person texts a male, then he-or-she may get a reply.

If a male looks around, then he may hope to see someone.

If a person doesn't like some particular store, then he-or-she may not keep going to it.

While some bad rules come from parsing or processing mistakes, these are less of a problem than the heavy tail of difficult constructions. For instance, there are idioms that we want to filter out (*e.g.*, 'I'm embarrassed but…') and, as in the early work of Hearst (1992), other bad outputs show context-dependent rather than general relations:

If a girl sits down in a common room, then she may hope to avoid some pointless conversations.

The sitting-down may not have been *because* she wanted to avoid conversation but because of something prior.

It's difficult to compare our results to other systems because of the differences of representation, types of rules, and evaluation methods. The best performing method from ISP (Pantel *et al.*, 2007), ISP.JIM, achieves 0.88 specificity (defined as a filter's probability of rejecting incorrect inferences) and 0.53 accuracy. While describing their Sherlock system, Schoenmackers *et al.* (2010) argue that 'the notion of "rule quality" is vague except in the context of an application' and thus they evaluate the Horn clauses they learn in the context of the Holmes inference-based QA system, finding that at precision 0.8 their rules allow the system to find twice as many correct facts.

5.1.4 *Conclusions*

Enabling an inference system to reason about common situations and activities requires more types of general world knowledge and lexical knowledge than are currently available or have been targeted by previous work. We've suggested an initial approach to acquiring rules describing complex consequences or reasons and subtle relations among

adjectival attributes: We find possible rules by looking at interesting discourse patterns and rewriting them as conditional expressions based on semantic patterns.

A natural question is why we don't use the machine-learning/bootstrapping techniques that are common in other work on acquiring rules. These techniques are particularly successful when (a) they are aimed at finding fixed types of relationships, such as hyponymy, near-synonymy, part-of, or causal relations between pairs of lexical items (often nominals or verbs); and (b) the fixed type of relationship between the lexical items is hinted at sufficiently often either by their co-occurrence in certain local lexico-syntactic patterns, or by their occurrences in similar sentential environments (distributional similarity). But in our case, (a) we are looking for a broad range of (more or less strong) consequence relationships, and (b) the relationships are between entire clauses, not lexical items. We are simply not likely to find multiple occurrences of the same pair of clauses in a variety of syntactic configurations, all indicating a consequence relation – you're unlikely to find multiple redundant patterns relating clauses, as in 'Went up to the door but didn't knock on it'.

5.2 Sharpening to Appropriate Form

Lore produces a large volume of factoids from text that express possibilistic general claims such as *A person may have a head* or *People may say something*. This section presents a rule-based method to sharpen certain classes of factoids (ILFs) into stronger, quantified claims such as *All or most persons have a head* or *All or most persons at least occasionally say something* – statements strong enough to be used for inference. The judgement of whether and how to sharpen a factoid depends on the semantic categories of the terms involved, and the strength of the quantifier depends on how strongly the subject is associated with what is what is predicated of it. This section concludes with an initial assessment of the quality of this automatic strengthening of knowledge and the next chapter demonstrates and evaluates the use of sharpened formulas for inference.

5.2.1 *From Weak Knowledge to Strong Knowledge*

Human-level artificial intelligence, as required for problems like natural language understanding, seems to depend on the ability to perform commonsense reasoning. This in turn requires the availability of considerable general world knowledge in a form suitable for inference. There are several approaches to acquiring such knowledge, including directly interpreting general statements such as glosses in dictionaries (*e.g.*, Clark *et al.*, 2008a), abstracting from clusters of propositions (Van Durme *et al.*, 2009), and the hand-authoring of rules, as in Cyc (Lenat, 1995). Hand-coding is apt to be haphazard in its relationship to language as it depends on the cerebration of numerous knowledge engineers with differing intuitions and no particular commitment to consistency with language or across domains. The volume of hand-coded knowledge produced so far is also probably a couple orders of magnitude short of what is needed.

The approach of this section is to begin with the large volume of weak, general factoids discovered by Lore, select factoids that lend themselves to logical strengthening, and then sharpen these into quantified general statements that can be used in combination with other facts to generate new conclusions by forward inference.

For example, the following is a factoid and its automatic English verbalization:

[⟨det elm_tree.n⟩ have.v ⟨det branch.n⟩]
An elm tree may have a branch

In this case, the factoid was extracted from text that referred to a 'branch of an elm tree'. A text occurrence like this presumes that *at least sometimes* an elm tree has a branch, so this is how we understand and verbalize the formula.

However, this claim is not as strong as we would like it to be: It can be strengthened to say that all – or at least most – elm trees have a (*i.e.*, at least one) branch and that having this branch is an episode that is generally permanent with respect to the tree's existence. Using quantifier restrictors, we represent this as

(all-or-most *x*: [*x* elm_tree.n]
(some *e*: [[*x* | *e*] permanent]
(some *y*: [*y* branch.n]
[[*x* have-as-part.v *y*] ∗∗ *e*])))

As the predicate *have* is used in many ways, we have disambiguated it here to *have-as-part*. Other senses of *have*, such as *give-birth-to* or *possess*, and other kinds of predicates will require different quantification.

In this section, we present a general, rule-based method of doing such quantificational sharpening using existing lexical semantic categories and corpus frequencies. We change unscoped quantifiers to scoped ones and estimate the frequency of events/times and subjects for which each factoid is likely to hold. We then show some simple examples of commonsense reasoning using multiple sharpened premises.

5.2.2 *Strengthening Factoids*

For much of Knext's output, the weak formulation is as much as we want to assert. So when sharpening, we want to focus on those factoids that are likely targets to be strengthened. The method for doing so is to write rules that match large sets of factoids to patterns using semantic predicates. A simplified example of a rule is

```
(/ ((det? animal?) have.v (det? animal-part?))
   (all-or-most x (x animal?)
      (some e ((pair x e) enduring)
         (some y (y animal-part?)
            ((x have-as-part.v y) ** e)))))
```

The first argument to / is the pattern to be matched, using functional predicates ending in question marks to check elements. If the input matches, the output form is instantiated with the matching parts of the ILF bound to the name of each predicate. More

complicated rules allow for repetition of arguments, alternatives, functional attachment in the output form, *etc.*

Without sharpening, Knext learns that *A person may have a head*, but we know that having a head isn't optional: it's a crucial part of being a (living) person. Even for body parts that can be lost, it's reasonable to conclude that *most* people have them, so this is what the rule asserts. (There are also ephemeral parts, such as a leaf on a tree, in which case it is inappropriate to say that having the part is permanent with respect to the tree's existence. Such cases are few and can be hand enumerated.) To identify an *animal-part* above, we make use of the nominal hierarchy from WordNet (Fellbaum, 1998), which classifies most of these as hyponyms of *body part*. Similar rules are used to match plants and artifacts with their respective parts.

Note that the *part-of* relations expressed in sharpened factoids needn't be in WordNet. For example, in a factoid of type [⟨det contraption.n⟩ have.v ⟨det button.n⟩], we would interpret this as a have-as-part.v sentence as long as WordNet treats *some* sense of 'button' as part of *something*, such as a shirt, doorbell, cellphone, *etc.* The same is true for factoids like [⟨det chicken.n⟩ have.v ⟨det feather.n⟩] or [⟨det rose-bush.n⟩ have.v ⟨det flower.n⟩].

The quantifiers used in sharpening are *all* (∀), *all-or-most*, *most*, *most-or-many*, *many*, *many-or-some*, and *some* (∃). *Much* is substituted for *many* for mass predicates like 'oil'. Temporal quantification can be over *frequent* or *occasional* episodes or just *some* possible episode. (This is refined further in §5.3.) Quantifier strength is decided by a combination of semantic rules and corpus statistics. Non-repeatable predicates such as being born or dying are usually true of all individuals of a class as is having a body part, *supra*. On the other hand, few repeatable actions are universal; most are done by a smaller number of individuals. Breathing and eating food are some obvious exceptions, though even *eating* isn't universal when it takes an argument: All people eat, but how many eat babka?

5.2.3 *Quantificational Disambiguation*

The choice of quantifier scopes in sharpening is not as hard as the general scoping problem: Lore's abstraction does not produce strongly quantified factoids like 'A doctor may live in every city', only ones containing weak indefinites. So, unlike the empirical quantifier scope disambiguation of Srinivasan & Yates (2009), our approach can rely simple semantic patterns to produce scoped quantifier configurations.

5.2.4 *Lexical Disambiguation*

The most frequent relation in Knext-extracted factoids is 'have', so disambiguation of this 'light' predicate, to the extent necessary for inference, is of particular interest. However, it is not clear that all word senses need to be disambiguated. We claim that verbal predicates aren't nearly as ambiguous as has generally been assumed; they're just semantically *general*.

As the criterion for whether disambiguation is necessary, we ask whether or not the entailments follow from the argument types. For example, it's not strictly necessary to disambiguate 'have' in 'have an accident' since the only possible entailments of this phrase in actual use are those for the *experience* sense. By contrast, it is important for us to be able to narrow the sense of 'have' to *eat* in *A person may have a lobster* if that (rather than a possessive sense) is the intended meaning. So the appropriate sharpening would be as follows (where e is the eating episode characterized by the sentence, with the characterization relation indicated by the episodic operator '$**$'):

[⟨det person.n⟩ have.v ⟨det lobster.n⟩]
(many x: [x person.n]
(some e (some y: [y lobster.n]
[[x eat.v y] $** e$])))

Note that *have* often simply serves as a kind of particle binding a *relational* noun to the subject, as when we say 'John has a sister' or 'John has a (certain) weight'. It seems

pointless to invent separate meanings of *have* for all these cases, such as *have-as-relative* or *have-as-weight*; these meanings are already inherent in the nominals themselves:

[⟨det male.n⟩ have.v ⟨det sister.n⟩]
(many *x*: [*x* male.n]
(some *e* (some *y*: [*y* female.n]
[[*x* (have-as sister.n) *y*] ⁎⁎*e*])))

[⟨det male.n⟩ have.v ⟨det weight.n⟩]
(many *x*: [*x* male.n]
(some *e*: [[*x*|*e*] permanent]
(some *y* [[*y* weight-of.n *x*] ⁎⁎*e*])))

These relational uses of 'have' are identified based on the semantic categories of the subject (*e.g.*, a *causal agent* or *social group*) and the object (*e.g.*, a hyponym of *relative*, *leader*, or *professional*) while most features like 'weight' are hyponyms of *attribute*.

In some cases, disambiguation is necessary but is difficult enough that we choose not to sharpen the factoid rather than risk doing so incorrectly. A particularly difficult class of factoids to sharpen are those involving prepositions, where we need to at least implicitly disambiguate different uses, *e.g.*, 'a man with one arm' *vs* 'a man with a cake'. To avoid bad sharpened output, we also need to check for nouns that don't mean much when standing alone, *e.g.*, 'front', 'thing', or 'issue'. We want to avoid sharpening factoids involving such terms, at least when they occur as the subject of sentences with no object.

5.2.5 *Sharpening with Stage-Level Predicates*

A key distinction for sharpening is between individual-level and stage-level predicates (Carlson, 1977a). *Individual-level predicates* endure over most of the existence of the individual they're predicated of while *stage-level predicates* describe dynamic goings-on or transient situations. While we want to quantify stage-level predicates over individuals

and episodes (a *stage* being a temporal slice of an individual), an individual-level predicate is just quantified over individuals.

We assume that if an entity has a capacity, it is exercised at least occasionally. Thus we sharpen factoids about abilities to stage-level quantification over episodes of performing them. Some factoids Lore learns explicitly state that an individual may be *able* to do something:

[⟨det female.n⟩ able.a (ka speak.v)]
A female may be able to speak

(The *ka* operator indicates a kind of action.) Factoids like this can indicate abilities that are specific to a few individuals – say, being able to ride a horse – rather than generally true as in the example above. But they can also indicate basic abilities: We rarely state that someone is 'able to' do a basic action like walking. Yet, if someone breaks their leg, we might say that they are 'able to walk (again)' and can produce an appropriate factoid.

Sharpened factoids about abilities are also formed from factoids about actions without *able.a*, such as [⟨det female.n⟩ swim.v]. As a stage-level predicate, *swim* will lead to quantification over episodes:

(many x: [x female.n]
(occasional e [[x swim.v] $**$ e]))

What we aim to get are formal versions of habitual sentences like

Most people occasionally use a cellphone.
Most companies occasionally announce a product.

rendered in the following manner:

(most x: [x person.n]
(occasional e (some y: [y cellphone.n]
[[x use.v y] $**$ e])))

(most *x*: [*x* company.n]
(occasional *e* (some *y*: [*y* product.n]
[[*x* announce.v *y*] ⁎⁎ *e*])))

Stage-level adjectives also get quantified over episodes:

[⟨det male.n⟩ hungry.a]
(all-or-most *x*: [*x* male.n]
(occasional *e* [[*x* hungry.a] ⁎⁎ *e*]))

Some 'have' propositions represent temporally quantified occurrences, *e.g.*, 'All or most persons occasionally have a thought/cold/shock/party…' We recognize such a use by a subject who is a *causal agent* and an object that is a *psychological feature*, *event*, or *state*.

[⟨det male.n⟩ have.v ⟨det thought.n⟩]
(all-or-most *x*: [*x* male.n]
(occasional *e* (some *y*: [*y* thought.n]
[[*x* experience.v *y*] ⁎⁎ *e*])))

It would be distressing if we gave a similar quantification for the stage-level verb *die*. For this reason, stage-level predicates are divided into repeatable and non-repeatable ones. The latter includes strict once-per-existence predicates like *die* and also 'pivotal' ones like *marry*. While marriage is repeatable, we don't want to claim it's a *frequent* action for an individual, no matter how heavily reported marriages are in text. It is also necessary to distinguish those predicates that are nonrepeatable with respect to their objects, *e.g.*, while one can kill multiple times, one can only be killed once. Nonrepeatable predicates generally fall into a small number of VerbNet (Kipper-Schuler, 2006) categories, which we supplement with the other terms from the corresponding WordNet synsets.

5.2.6 *Sharpening with Individual-Level Predicates*

One class of factoids that should be strengthened to individual-level quantification are nouns predicated of noun phrases, such as [⟨det person.n⟩ fighter.n]. For these, the predicate in object position is most likely true of the individual's whole existence, *e.g.*,

(many-or-some x: [x person.n]
(some e: [[x | e] permanent]
[[x fighter.n] $**$ e]))

We avoid quantifying strongly when the complement is a role noun, such as those found in WordNet as hyponyms of *person* and ones like *weapon*: While it is true that a person can be a doctor, we don't want to conclude *most* people are doctors. And while a cane can be weapon (when it's used as such), we don't want to assert that most canes *are* (enduringly) weapons.

Most stative verbs are individual-level, but there are exceptions such as *own* or *contain*. This can depend on the subject: Books contain information for the duration of their existence, but a jug only contains water at specific times. Many people *at some time* own a house, but it is not necessary that they own them for their entire lives. Note that many verbs have both stative and dynamic senses, *e.g.*, *think* is stative in 'I think she's tired' but not in 'Think about the problem'. In sharpening factoids we don't generally attempt to distinguish between such uses – we allow a verb to be strengthened as a stative (and thus probably individual-level) if it has any stative sense.

We assume that a factoid expresses possession if the verb is *have* or *own*, there is an animate subject (such as a person or organization – hyponyms of *causal agent*), and the object is an artifact or a domesticated animal:

[⟨det person.n⟩ have.v ⟨det dog.n⟩]
(many x: [x person.n]
(some e **(**some y: [y dog.n]
[[x possess.v y] $**$ e]**)))**

Individual-level adjectives can be found by looking at the hypernyms of the derivationally related form in WordNet, so for *fond* we get *fondness*, which has *attribute* as a hypernym. This, *tendency*, and *quality* are good indicators of an individual-level property, *e.g.*,

[(k (plur cat.n)) fond.a (of.p (k milk.n))]
(most *x*: [*x* cat.n]
[*x* fond.a (of.p (k milk.n))])

5.2.7 *Sharpening with Kind-Level Predicates*

The above is a factoid about kinds (indicated by the *k* operator), which we sharpen to be about individuals of the kind. An additional type of factoid we haven't dealt with here is the type involving kind-level predicates, which are predicated not of individuals but of a whole genus, *e.g.*,

[(k (plur cow.n)) widespread.a]

A problem here concerns the level of abstraction: If we view this factoid as a statement about an individual, *viz.*, the kind *cows* (much as in 'The Milky Way is widespread'), we should not read it possibilistically as 'Cows may be widespread', but simply as 'Cows are widespread'. But when we abstract from the particular kind to species, we want to conclude that *some species* are widespread.

5.2.8 *Sharpening with Events*

Another special case are factoids with event nominal (*e.g.*, a war or a party) subjects, for which neither stage-level nor individual-level predicates should result in quantification over episodes. We identify these event nominals by using the verbalizations in the Nomlex nominalization lexicon and those words categorized as hyponyms of *event* and related synsets in WordNet.

It's worth noting that while we use WordNet as our primary resource for semantically categorizing predicates in the process of sharpening, our factoids express information beyond what's in WN: While it tells us that *writing* is a *human activity*, it does not tell us that people write *letters* and so on. It is only the combination of Knext factoids with WordNet, VerbNet, and other resources of lexical semantic features that provides the bulk of the sharpened output.

5.2.9 *Frequencies and Quantifier Strength*

We want to conclude that most men have shoes, but few men have a yacht. General sharpening rules are nudged to stronger or weaker quantification based on the strength of the association between the subject of the factoid and what is predicated of it – the normalized *pointwise mutual information* (PMI) computed over the entire KB of factoids being sharpened.[5] Taking our formulas as formal generic statements, this approach reflects the inductive view of generalizations: After we observe enough people possessing dogs (from textual references), we take it to be likely. And while Lore may learn only a few factoids about, say, Komodo dragons, if there's a high PMI between *Komodo dragon* and *eat carrion*, then it will be quantified as being true of most Komodo dragons. For discussion of whether generic statements (such as the commonsense knowledge we are trying to abstract from possibilistic factoids) should be understood inductively or from a rules-and-regulations view not dependent on real-world activity, see Carlson (1995).

5.2.10 *Evaluation of Sharpening*

As an initial evaluation of our sharpening methods, we first took a set of propositions extracted from the British National Corpus that were previously evaluated by crowd-

5 Mutual information measures have been used by others in knowledge-extraction, *e.g.*, Clark & Harrison (2009a) assign the strength – or plausibility – of DART's Knext-like extractions based on how mutual information to reflect how much the observed frequency reflects a true association between elements in a tuple.

sourcing on Mechanical Turk (see Appendix A). Non-experts were shown the English verbalizations of factoids (*e.g.*, *A man may have a head*) and asked to rate how well they conveyed accurate commonsense knowledge. Out of 1500 randomly sampled BNC factoids, 435 of them could be sharpened. The smaller size of this set represents a preference for precision over recall and the large number of factoids that don't seem to merit a stronger form, even among those that were judged to hold in when stated weakly. (Nonetheless, later work has increased the coverage of factoids that are sharpened.)

Here we want to judge whether the sharpened forms express reasonable general claims and have been strengthened sufficiently. The authors therefore judged 200 sharpened factoids on the same scale of 1–5 (with 1 being best) based on their agreement with the following primary and secondary statements:

Statement 1. The factoid is a reasonable general claim about the world even if – perhaps – it isn't as strongly quantified as it might be.

If so (that is, if the judge rates the factoid 1 or 2), they then judged

Statement 2. The quantifiers seem sufficiently strong.

So, for instance,

(some x: [x male.n]
(some e: [[x | e] permanent]
(some y: [y head.n y]
[[x have-as-part.v y] $**$ e]))))

would not be rated very well for Statement 2. It is true, but the claim should be quantified more strongly: All men have heads.

Since it is quite hard to produce a good sharpened statement from a bad factoid, we are interested not just in the overall performance of the sharpening but also in how it does on a subset of good factoids. For this, we took those factoids with an average Turker-assigned rating between 1 and 2. Such a rating means that, in its weak, possibilistic form, the factoid is probably a reasonable claim about the world.

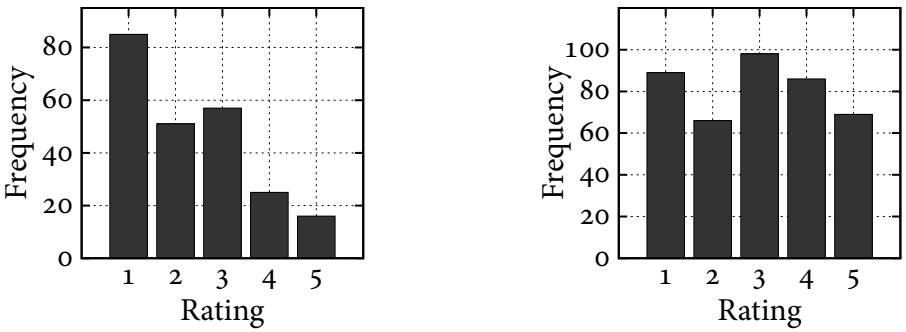

Figure 5.2: **Sharpened formula quality.** Agreement with Statement 1 for formulas sharpened from factoids rated 1–2 (left) and from all factoids (right). The vertical axis shows the number of factoids that were given each rating, counting both judges' responses. Lower ratings are better.

	Stmt 1 Avg.	*Stmt 2 Avg.*
Judge 1	3.01	1.66
Judge 2	2.73	1.71
Correlation	0.79	0.75

Table 5.2: **Ratings of sharpened factoids.** Includes those produced from all unsharpened factoids, not just the highly rated subset.

As seen in Figure 5.2 (left), these favorably judged weak factoids yield favorably judged strengthened factoids (when they yield any at all) more often than they yield ambivalent or negative judgements. While 36% of the ratings of factoids sharpened from the good unsharpened factoids (those with an average rating between 1 and 2) are rated a 1, only 21% of the complete set were so rated. As can be seen from the right histogram, judgements of sharpened factoids are considerably worse if the unsharpened factoids include everything generated. Therefore it is crucial to pre-filter unsharpened factoids, perhaps using crowdsourcing (as was done here) or by improved automatic methods. This can also include improvements in the initial knowledge extraction and in the technology it relies on. Incorrect syntactic parses, including improbable parts-of-speech, were evident in the judged results: Any improvements in parsing are likely to also improve our knowledge extraction.

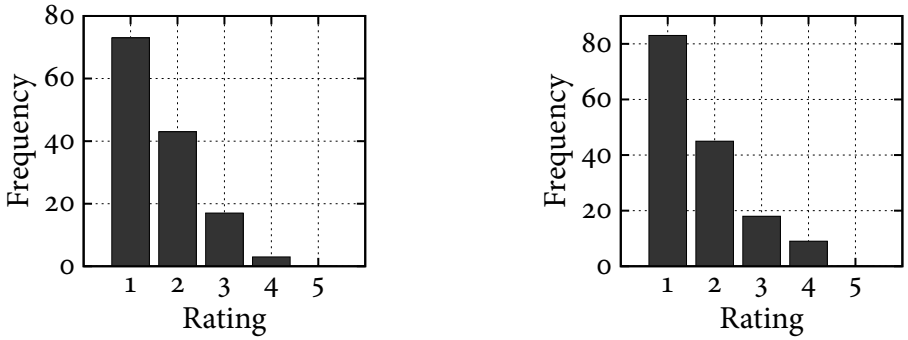

Figure 5.3: **Sharpened formula quantifier strength.** Agreement with Statement 2 for formulas sharpened from factoids rated 1–2 (left) and from all factoids (right). The vertical axis shows the number of factoids that were given each rating, counting both judges' responses. Lower ratings are better.

5.2.11 *Conclusions*

In this section we have suggested that we can sharpen classes of factoids expressing commonsense knowledge from weak, possibilistic claims to stronger, quantified claims. To do this we made use of semantic categories for disambiguating predicates and recognizing factoids that express characterizing properties that deserve strengthening. We also made use of the frequency with which we extract claims from text to induce the strength of that quantification. Initial evaluation suggests that the resulting strengthened factoids are of good quality, though improvement is needed for them to be suitable for inference.

5.3 Learning Expected Event Frequencies

Commonsense reasoning requires knowledge about the frequency with which ordinary events and activities occur: How often do people eat a sandwich, go to sleep, write a book, or get married? This section presents work to acquire a knowledge base pairing factoids about such events with frequency categories learned from simple textual patterns. A collection of factoids with the resulting event frequencies are evaluated for accuracy, and I demonstrate the application of the results to the problem of knowledge refinement as discussed in the previous section.

The previous section used sharpening rules that distinguished only three types of temporal predications:

1 those that hold for the existence of the subject (*individual-level*), *e.g.*, a house being big;
2 those that hold at a specific moment in time (*non-repeatable stage-level*), *e.g.*, a person dying; and
3 those that hold at multiple moments in time (*repeatable stage-level*), *e.g.*, a person drinking a cup of coffee, which are quantified as 'frequent' or 'occasional' events based on the association between subject and predication

However, repeatable stage-level predications vary from those done with great frequency, such as a person saying something, to those done quite infrequently, such as a woman giving birth. This section describes a simple method to learn rough frequencies of such events from text.

Our focus is on the commonsense knowledge needed for many AI applications, rather than more specific domain knowledge, so we look for the frequency of everyday events – such as going to work – that might be mentioned in ordinary text like newspaper articles, rather than big events – like earthquakes devastating a city, which tend to be rare and unpredictable – or small events – like atoms decaying, which would typically escape our notice.

5.3.1 *Previous Work*

We are unaware of any previous work aimed at systematically learning the expected or normal frequency of events in the world. However, our basic approach to this problem aligns with a long-running line of work using textual references to learn specific kinds of world knowledge, which has been popular at least since Hearst (1992) used lexico-

syntactic patterns like 'NP_0 such as {NP_1, NP_2, …, (and|or)} NP_n' to learn hyponym relations, *e.g.*, 'Bambara ndang is a bow lute' from large text corpora.[6]

In addressing the problem of *quantificational* disambiguation, Srinivasan & Yates (2009) learn the expected sizes of sets of entities that participate in a relation; *e.g.*, how many capitals a country has or how many cities a person tends to live in. They do this by using buckets of numeric phrases in hand-crafted extraction patterns like '(I|he|she) ⟨word⟩+ ⟨numeric⟩ ⟨noun⟩', which would match 'she visited four countries'. They apply these patterns to Google's Web 1T n-gram Corpus.

Gusev *et al.* (2011) presented a similar approach to learning event durations using query patterns sent to a Web search engine, *e.g.*, '$\langle \text{event}_{\text{past}}$ for * ⟨bucket⟩', where the bucket is a category in [seconds, minutes, hours, …, decades] for classifying the event's expected duration. Both of these papers are notable for gaining wide coverage by indirectly using Web-scale text. However, they are limited by the brevity of patterns in n-grams and by the coarse matching abilities of Web queries, respectively. §5.3.3 discusses these trade-offs and our approach, focusing on large offline corpora.

The contribution of this section is the application of a traditional technique to a new problem. Temporal frequencies are of key importance to improving the quality of automatically learned knowledge for commonsense reasoning. Additionally, we hope that providing a knowledge base of expected frequencies for factoids about everyday events will serve as a new resource for other work in knowledge extraction and reasoning.

5.3.2 *Textual Patterns of Frequency*

The most direct linguistic expression of temporal frequency comes from frequency adverbs: words like *usually* and *always*, distinct in their meaning from other adverbs of quantification like *twice*. Sentences that contain a frequency adverb are referred to as *frequency statements*, *e.g.*, 'John sometimes jogs in the park.' Frequency statements are

6 This example has probably provided more press for the bambara ndang in information-extraction than in all of musicology or anthropology.

interesting because their truth depends not just on the existence of some past events that support them but on a regular distribution of events in time. That is, saying that John 'sometimes jogs' means that it is a habitual rather than incidental activity.

As Cohen (1999) observes, much of our knowledge about the world is expressed through frequency statements, but it's not entirely clear what these sentences mean. From the perspective of knowledge extraction, they can seem quite opaque as their meaning seems to rely on our pre-existing ideas of what a normal temporal frequency for the event would be. For instance, to say that 'Mary snacks constantly' (or 'frequently' or 'occasionally') only makes sense if you already have in mind some range of frequencies that would be normal or unremarkable.

More absolute frequency adverbials, such as *daily*, *weekly*, or *every other week* avoid the problem of depending on a person's expectations for their meaning. However, these tend to occur with extraordinary rather than ordinary claims. For instance, in the British National Corpus we find

> Clashes between security forces and students had occurred almost daily.
>
> New [viruses] are discovered every week.

Both of these are expressing surprising, unexpected information.

Following the example of §5.1 in considering 'disconfirmed expectations', we look for textual expressions that indicate a person's frequency expectation has not been met and, looking at these in aggregate, we conclude what the original, implicit expectation is likely to have been. An example of such a disconfirmed expectation is

> Bob hasn't slept in two days.

The production of sentences like this suggests that this is an unusually long gap between sleep periods for most people. We are unlikely to find many sentences saying, *e.g.*, 'Bob hasn't slept in two hours' as this would not defy our expectation. (And while we will find exaggerations, such as 'I hadn't slept in weeks', the classification technique we describe will favor the smaller interval unless the counts for a longer interval are quite high.)

In this initial approach, we make use of two other patterns indicating temporal frequency. An additional indication of an upper-bound on how infrequent an event tends to be is a reference to the last time it was completed, or the next time it's anticipated, *e.g.*, 'He walked the dog yesterday' or 'She'll go to the dentist next month'.

The other pattern is the use of *hourly*, *daily*, *every week*, *etc.* While frequency statements with such adverbs can be communicating a frequency that's much higher or lower than expected, they serve as an important source of information when we don't find matches for the defied expectations. They also occur as prenominal modifiers: For a factoid like *A person may eat bread*, we want to match references to 'his daily bread'. This use is presumptive and, as such, indicates a usual or expected frequency, as in 'our weekly meeting' or 'the annual conference'.

5.3.3 *Method*

Rather than relying on query-based retrieval from the Web, or on the use of n-gram databases, we have chosen to process a selection of large text corpora including the Brown Corpus (Kučera & Francis, 1967), the British National Corpus (BNC Consortium, 2001), the Penn TreeBank (Marcus *et al.*, 1994), Gigaword (Graff *et al.*, 2007), a snapshot of English Wikipedia (Wikipedia, 2009), a collection of weblog entries (Burton *et al.*, 2009), and Project Gutenberg e-books (Hart & volunteers, 2006).

The motivation for doing so is the larger context offered and the flexibility of matching. Search engine queries for patterns are limited to quoted strings, possibly containing wildcards: There's no reasonable mechanism to prevent matching patterns nested in a sentence in an unintended way. For instance, searching for 'I hadn't eaten for months' can easily match not just the expected hyperbole but also sentences like 'I felt like I hadn't eaten for months'. Sets of n-grams pose the problem of limiting pattern length. While it's possible to chain n-grams for longer matches, this forfeits the guarantee of any actual sentence containing the match.

As a set of appropriate, everyday events abstracted away from specific instances, we used a corpus of factoids (ILFs) learned most frequently by Lore. We heavily filtered the knowledge base both for quality (*e.g.*, by limiting predicate names to known words) and to focus on those factoids describing the sort of action to which we want to assign a frequency. This included removing passives ('A person may be attacked') and subjects that aren't causal agents (according to WordNet). We abstracted multiple subjects to low common hypernyms for compactness and to focus on classes of related individuals, such as 'a parent', 'an executive', or 'a scholar'.

A good indication that a factoid can be annotated with a frequency is telicity: Telic verb phrases describe events rather than continuous actions or states. To check if the predication in a factoid is possibly telic, we look in the Google n-gram data set for short patterns. For each factoid of form (*x y z**) and each set of indicators *s*,

(quickly|immediately|promptly)
(suddenly|abruptly|unexpectedly)
(inadvertently|unintentionally|deliberately|unwittingly|purposely|accidentally)
(repeatedly|frequently)

we look for: '*s x y*ed *z**', '*x y*ed *z* s*', and '*x s y*ed *z**' where *x* is the subject, *y*ed is the past tense of the verb, and *z* consists of any arguments. Any factoid with non-zero counts for more than one set of indicators was considered possibly telic. For each possibly telic factoid, we first determine whether it describes a regular event or not. A regular event doesn't need to be a rigid, scheduled appointment, just something done fairly consistently. 'Brush your teeth' is regular, while 'Overcome adversity' is not; it depends on some scenario arising. Regularity can be indicated explicitly:

*y*s/*y*ed regularly/habitually
*y*s/*y*ed invariably/inveterately/unvaryingly
*y*s/*y*ed like clockwork

*y*s/*y*ed at regular intervals

It can also be suggested by a stated interval:

*y*s/*y*ed hourly/daily/weekly/monthly/yearly/annually
*y*s/*y*ed every hour/day/week/month/year
every hour/day/week/month/year *x y*s/*y*ed

If we don't match enough of these patterns, we don't consider the factoid to be regular; It may be an occasional or existence-level predication, or we may just lack sufficient data to determine that it's regular.

For each regular-frequency factoid, we then check the corpora for matches in our three categories of patterns:

Explicit Frequency Matches These indicate the exact frequency but may be hyperbolic. The 'hourly' and 'every hour' style patterns used for checking regularity are explicit frequency indicators. In addition, if the factoid contains 'may have a *z*', we search for the prenominal modifiers:

's/his/her/my/your/our hourly/daily/weekly/monthly/yearly/annual *x*

Disconfirmed Expectation Matches These indicate that people expect the activity to be done 'at least *bucket* often'. These include many small variations along these lines:

Hourly/multiple times a day:
Has *x y*ed this morning/afternoon/evening?
Didn't *x y* this/last/yesterday morning/afternoon/evening?
Hasn't *y*ed for/in over an hour
Has not *y*ed for the whole/entire day

Daily/multiple times a week:
Have *x* not *y*ed today?
Did *x* not *y* today/yesterday?

Had not *y*ed for/in more than *n* days

Haven't *y*ed for the whole/entire week

Weekly/multiple times a month:

Haven't *x y*ed this week?

Didn't *x y* this/last week?

Hadn't *y*ed for more than a week

Had not *y*ed for the whole/entire month

Monthly/multiple times a year:

Hasn't *x y*ed this month?

Did *x y* this/last month?

Hadn't *y*ed for over *n* months

Hadn't *y*ed for the whole/entire year

Yearly/multiple times a decade:

Have *x y*ed this year?

Didn't *x y* this/last year?

Haven't *y*ed for/in over a year

Hadn't *y*ed for an entire decade

Last Reported Matches These are statements of the last time the predication is reported as being done or when it's expected to happen next. These are useful, as you wouldn't say 'I took a shower last year' if you take one daily. They indicate that the event happens 'at most *bucket* often'.

Hourly/multiple times a day:

*y*ed an hour ago

*y*ed earlier today

'll/will *y* later today

Daily/multiple times a week:

*y*ed today/yesterday

*y*ed on Sunday/…/Saturday

'll/will *y* tomorrow/Sunday/…/Saturday

'll/will *y* on Sunday/…/Saturday

Weekly/multiple times a month:

*y*ed this/last week(end)

'll/will *y* next week(end)

Monthly/multiple times a year:

*y*ed this/last month

'll/will *x* next month

Yearly/multiple times a decade:

*y*ed this/last year/season/spring/…/winter/January/…/December

'll/will *y* next year/season/spring/…/winter/January/…/December

Decision For each of the three categories of patterns, we select the frequency bucket that it most strongly supports: We iterate through them from *hourly* to *yearly*, moving to the next bucket if its count is at least 2/3 that of the current one. For the 'last reported' matches, we go in the opposite direction: *yearly* to *hourly*.

From the three choices, the two buckets with the highest supporting counts are selected. If the range of these buckets is wide (that is, there is more than one intervening bucket), the bucket for a more frequent reading is chosen; otherwise, the less frequent one is chosen. This choice compensates for some hyperbole: If people claim they haven't slept for *days* and for *years*, we choose *days*. However, if we find that people haven't showered for *hours* or *days*, we choose days as a reasonable lower bound.

5.3.4 *Evaluation*

To evaluate how accurately this method assigns an expected frequency to a factoid, we sample 200 factoids that were classified as describing a regular occurrence. Each of these is verbalized as a conditional, *e.g.*,

> *If a person drives taxis regularly, he or she is apt to do so daily or multiple times a week.*
>
> *If a male plays (video games) regularly, he is apt to do so daily or multiple times a week.*

Note that we do not take the factoid to apply to all possible subjects, but for those it applies to, we're indicating our expected frequency. Arguments are taken to be narrow-scope, *e.g.*, for 'a person may greet a friend', it can be a different friend for each greeting event rather than the same friend every time.

For each of the sampled factoids, two judges evaluated the statement 'This is a reasonable and appropriately strong frequency claim (at least on some plausible understanding of it, if ambiguous)' on this scale:

1 Agree
2 Lean towards agreement
3 Unsure
4 Lean towards disagreement
5 Disagree

The average rating for Judge 1 was 2.45, the average rating for Judge 2 was 2.46, and the Pearson correlation was 0.59.

A simple baseline for comparison is to assign the most common frequency ('daily') to every factoid. However, for this to be a fair baseline, this needs to be done at least for the entire possibly-telic KB, not just the factoids identified as being regular, as that classification part of the method being evaluated. This baseline was evaluated for 100

factoids, with an average ratings of 3.06 and 3.51 (correl. 0.66) – worse than 'unsure'. This result would be even lower if we applied this frequency to all factoids rather than just the telic ones: We would claim, for instance, that a person has a head daily.

The authors also judged a random sample of 100 of the factoids that were marked as not being regular actions. These were verbalized as denials of regularity:

Even if a person files lawsuits at all, he or she doesn't do so regularly.

Of these, on average the judges indicated that 30 could reasonably be thought to be regular events that we would like to assign a frequency to.

These annotations serve as a guide in the sharpening of Lore factoids into full Episodic Logic forms. For instance, from the factoid *A person may eat lunch*, we can now select the correct episodic quantifier *daily*:

(all-or-most *x*: [*x* person.n]
  (daily *e*
    (some *y*: [*y* lunch.n]
      [[*x* eat.v *y*] ** *e*])))

That is, for all or most persons, there is a daily episode that is characterized by the person eating some lunch.

5.3.5 *Conclusions & Future Work*

The acquisition of temporal frequency information for everyday actions and events is a key problem for improving automatically extracted commonsense knowledge for use in reasoning. We argue that this information is readily available in text by looking at patterns expressing that a specific instance is at odds with the expected frequency, those that report frequencies explicitly, and those stating the last time such an event occurred. We find that a simple approach assigns event frequencies with good accuracy, allowing us to improve the temporal quantification of knowledge learned by Lore.

There is room to improve the frequency labeling, for instance, using machine-learning techniques to combat sparsity issues by discovering new textual patterns for event frequencies. It would also be interesting to see how performance could be improved by automatically weighting the different patterns we've discussed as classification features.

5.4 Chapter Summary

This chapter presented Lore, a system for learning commonsense knowledge from text. While previous work on Knext has found a great volume and variety of possibilistic general world knowledge, this lacks much of what we consider common sense, such as the expected outcome of actions. Lore expands Knext's extractions, focusing on patterns such as 'disconfirmed expectations' that reveal what people expect to normally be true – an approach to overcoming the reporting bias of text discussed in Chapter 4. To provide appropriately strong, partially disambiguated knowledge that can be used for inference, Lore sharpens the initial logical forms learned from text, using lexico-syntactic rules and corpus frequencies. A variety of text patterns were demonstrated to find the expected frequencies of events to allow more specific temporal quantification in sharpening. In the next chapter we demonstrate and evaluate the use of Lore's output for inference.

6 Making & Evaluating Inferences with Commonsense Rules

> Most of their remarks were the sort it would not be easy to disagree with: 'What I always say is, when a chap's hungry, he likes some victuals,' or 'Getting dark now; always does at night,' or even, 'Ah, you've come over the water. Powerful wet stuff, ain't it?'
>
> C. S. LEWIS, *The Voyage of the Dawn Treader*, 1952

Human-level artificial intelligence requires the ability to reason about the sorts of everyday situations and individuals that people write about. Consistent with the view that language is a mirror of mind, in this dissertation we acquire general commonsense factoids from predication and modification structures in text. Then, using existing lexico-semantic resources and textual frequencies, we sharpen that knowledge into quantified axioms, expressed in a natural language–like logic. This chapter demonstrates uncertain inference with these rules and evaluates the reasonableness of the conclusions, comparing their quality to that of conclusions drawn from large knowledge bases without sharpening.

6.1 Evaluating Knowledge Extraction

There are many ways of evaluating a collection of knowledge. An important measure is quantity: The more knowledge available for reasoning, the more useful a knowledge base may be. However, not all accurate knowledge is equally informative. While it is true that *Sparrows have feathers*, *Nightingales have feathers*, and *Ravens have feathers*, and so on, a

knowledge base containing all of these facts is arguably less useful than one that simply contains the knowledge that *Birds have feathers* (and that members of these species are birds). That is, knowledge that is entailed by a smaller KB is superfluous. However, storing knowledge at a greater generality than is justified will lead to inaccurate conclusions.

It is difficult to apply a notion of *recall* to general knowledge extraction since it is difficult to know how much – let alone, what – good knowledge you have *not* learned from a particular source. A labor-intensive approximation would have multiple experts read a set of documents and enumerate all the knowledge suggested in each document. Then you could measure how much of the knowledge identified by multiple readers is found by a KE system. Any knowledge found from these texts at an incorrect level of generality or not listed by the experts would then count against *precision*. The creation of such a data set has not been attempted, due to the effort, cost, and difficulty of agreement.

An even harder question in evaluating a knowledge acquisition effort is its *coverage* of the notional collection of all commonsense knowledge that would be useful for generally intelligent reasoners. While the core of commonsense knowledge is largely unchanging, there is an amorphous boundary between common sense and useful general world knowledge, which changes and grows over time. No approach has been suggested that would estimate how close we are to this ambiguous goal.

6.2 Knowledge for Reasoning

Many of the hardest problems of artificial intelligence seem to require commonsense knowledge and the ability to use it to reason about specific situations. For instance, creating an agent that can converse about everyday topics seriously (rather than with the superficiality of a 'chatbot') requires knowing about the things people talk about: other people, work, pets, politics, and much more. When someone mentions 'my friend Molly', you recognize 'Molly' as a female name, and you have a variety of expectations about friends. An intelligent agent needs similar knowledge, and the ability to draw commonsensical conclusions is an important evaluation of a collection of inferential knowledge.

Quantifier	*Weight*
All-or-most	0.8
Most	0.7
Most-or-many	0.6
Many	0.5
Many-or-some	0.4

Table 6.1: **Quantifier conclusion weights.**

McCarthy (1959) wrote, 'A program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows.'

The inference engine for Episodic Logic is Epilog (Schaeffer *et al.*, 1993), recently re-implemented as Epilog 2. EL and Epilog have proved their versatility and effectiveness in experimental applications ranging from processing aircraft maintenance reports to reasoning about fairy-tales. While these applications have been on a small scale, Epilog has also been shown to hold its own against state-of-the-art systems in the area of shallow theorem-proving over significantly large first-order knowledge bases (Morbini & Schubert, 2009), despite the fact that it handles a much richer representation than FOL.

Epilog supports input-driven *forward inference*, generating conclusions by combining newly obtained facts with existing knowledge, and goal-driven *backward inference*, working backwards from consequent to antecedent to check if there is evidence to support a specified conclusion. To enable inference with the generalized quantifiers used in this dissertation, we use specialized *modus ponens*–like axiom schemas, *e.g.*, for 'many',

$$(\forall_{\text{pred}}\, p, q\ (\forall_{\text{term}}\, b\ [[[b\ p]\ \text{and}\ (\text{many}\ x\text{: }[x\ p]\ [x\ q])]$$
$$\Rightarrow ((\text{adv-s}\ (\text{with-certainty}\ .5))\ [b\ q])))$$

With this rule, if we know that [John.name male.n] and (many x: [x male.n] [x man.n]), we can draw the very natural conclusion that

Certainty	*Verbalization*
0.9	Almost certainly
0.8	Very probably
0.6	Probably
0.4	Quite possibly
0.1	Possibly

Table 6.2: **Verbalizations of certainty values associated with statements.** Certainties can be assigned either as the result of inference or by a default axiom.

((adv-s (with-certainty .5)) [John.name man.n])

John is quite possibly a man.

The correspondence between numeric certainties of formulas and their verbalizations is given in Table 6.2.

Note that this rule could be stronger – *most males are men* (taking 'male' in the sense of male human and 'man' in the sense of male adult). However, it's important that the commonsense knowledge we learn from text not be too strongly quantified as we want explicitly stated facts – user input or assertions from a text being read – to outweigh it. To give default weights to explicitly stated knowledge, we introduce a meta-axiom that all well-formed formulas ϕ that don't begin with a certainty modifier imply ((adv-s (with-certainty .9)) ϕ).[1] Then, to resolve possibly conflicting conclusions, we can compare the certainty scores for ϕ and $\neg\phi$.

To see why this is necessary, consider the statement 'John is a nurse'. This is interpreted as [John.name nurse.n] and also, by the name matching a gazetteer[2], [John.name male.n]. From other text, we may have learned the unreliable rule that *Most nurses are female*. When we query the conclusion that [John.name female.n], Epilog will prove that

1 In future, this certainty value could be varied depending on the credibility of the source of the knowledge.

2 Gazetteers – lists abstracting names to classes of individuals – are used, both in knowledge acquisition and in interpretation of sentences. These include US presidents, cities, rock stars, corporations, tycoons, and more.

he is with a certainty of .7. However, when we query (not [John.name female.n]), Epilog will prove that he is not female with a certainty of .9, making that the favored conclusion.

The additional necessary axioms that (all *x*: [*x* male.n] (not [*x* female.n])) and vice versa are generated from the antonym relations from WordNet (Fellbaum, 1998). Although these don't cover all contradictory conclusions, *e.g.*, that a cat cannot also be a dog, they provide a number of important lexical relations.

6.2.1 *Uncertain Quantifier Chaining*

Furthermore, we want to sanction some rough-and-ready inferences involving quantifier chaining. For instance, if we know the rules

(all-or-most *x*: [*x* lion.n]
[*x* predator.n])
All or most lions are predators.

(many *x*: [*x* predator.n]
(some *e*: [[*x*|*e*] enduring]
[[*x* violent.a] ** e]))
Many predators are violent.

Then we can multiply their quantifier weights (Table 6.1) and form the rule that *Many or some lions are violent*. This is expected to hold in the absence of further information, *i.e.*, when we have no reason to suppose that lions are exceptional as predators with respect to the 'violent' property. So, any given lion is quite possibly violent.

Note that in such inferences, the quantifier in the conclusion will in general be weaker than the quantifiers in the premises (except that *all* maintains the full strength of whatever the other quantifier expresses). We make the slightly unusual assumption that quantifiers like *all-or-most*, *all*, and *most* have 'existential import'. That is, unlike in FOL, the quantifiers imply *some*. We even make the stronger assumption that these quantifiers

imply *many*. For example, if we are given that *All or most elm trees have branches*, we'll conclude that *Many elm trees have branches*. (This simply is an assumption that all the types that occur in our formulas have numerous instances, whether they be people, dogs, times, legs, *etc.*)

Note that these are being interpreted as *proportional* quantifiers, not absolute ones. The quantifier *all* of course is also proportional, and means 100%. *Some* is not treated as proportional; it just means 'at least one'. Temporal quantifiers such as *occasional* are not proportional either, but are stronger than *some*. For example, if Kevin occasionally smokes, and when he smokes he occasionally coughs, then he occasionally (not just at least once!) coughs, even though it will be at a lower frequency than he smokes.

6.3 Inferential Evaluation

In evaluating the inferences that are enabled by a KB, it is natural to consider a task-based evaluation such as demonstrating improvement at question answering. However, Kaplan & Schubert (2001) observed that

> …the task-based approach to evaluation…would mean giving up on one of the main attractions of the symbolic approach to AI, namely the idea that a system's internal representations can be interpretable by a human.

Even as more appropriately commonsense-oriented tasks have been presented, such as the Choice of Plausible Alternatives (Roemmele *et al.*, 2011), task-based evaluations continue to require a significant investment of time into system-building, which can be premature given the work yet to do on knowledge acquisition. It can also be quite difficult to measure the impact of a KB as performance on an end task often depends on multiple factors within the overall system.

For instance, in recent years, inferential knowledge has been applied to *recognizing textual entailment* (RTE). This is the task of judging for pairs of sentences whether the first (*t*) entails the second (*h*), where *entailment* is a semantic relation that holds only if

h is true in every possible world where *t* is true (Chierchia & McConnell-Ginet, 2000). A somewhat logical–inferential approach to RTE has been pursed by Clark & Harrison (2009b) among others. However, even given appropriate knowledge, using inference to improve RTE is currently limited by our ability to accurately interpret arbitrary premise sentences to trigger inference. Such open-ended semantic interpretation is an important and difficult area of research that requires more work, which is outside the practical scope of this dissertation.

Considering the evaluation of entailment rules (a subset of inference rules), Szpektor *et al.* (2007) argued:

> While measuring the impact of learned rules on applications is highly important, it cannot serve as the primary approach for evaluating acquisition algorithms for several reasons. First, developers of acquisition algorithms often do not have access to the different applications that will later use the learned rules as generic modules. Second, the learned rules may affect individual systems differently, thus making observations that are based on different systems incomparable. Third, within a complex system it is difficult to assess the exact quality of entailment rules independently of effects of other system components.

In this work we continue in the dominant evaluation strategy for knowledge extraction work of relying on human judgement (*e.g.*, Lin & Pantel, 2001; Barzilay & Lee, 2003; Sekine, 2005; Akbik & Löser, 2012). As observed in §5.1, judging inferential knowledge out of context can result in low inter-annotator agreement. This problem was previously identified by Szpektor *et al.* (2007) for the evaluation of entailment rules, leading them to propose instance-based evaluation. They presented judges with an entailment rule and a sample of sentences that match its antecedent for which they are asked whether the consequent holds. For samples of output from DIRT (Lin & Pantel, 2001) and TEASE (Szpektor *et al.*, 2004), they found this approach gave an improvement in agreement.

We want to judge the 'reasonableness' and the appropriate strength of uncertain conclusions in order to evaluate the quality of the axioms that have been extracted and the benefit that the sharpening of formulas gives to our ability to make commonsense inferences. To this end, we introduce fictional, named individuals of a variety of classes (US presidents, scientists, writers, cars, singers, artists, world cities, rivers, dictators, and countries) and – in the absence of other, specific information – we draw conclusions about them. For evaluation we rely on fictional instances rather than known individuals to avoid spurious ratings based on idiosyncrasies. For instance, it is a good commonsense claim that US presidents are elected to that office, but this would be an incorrect conclusion about Gerald Ford.

6.3.1 *Evaluating Factoids and Sharpened Formulas*

In addition to applying the sharpened rules, we want to draw conclusions based on unsharpened factoids to serve as a baseline. While the quantificational structure of the sharpened rules determines which can be applied, in selecting factoids to be used for pseudo-inference, we match only those with the predicate of interest in the leftmost unscoped quantifier. For instance, for 'us-president', we select *A US president may have an administration*, but not *A person may vote for a US president*. While the latter is perfectly reasonable, and we know that Barack Obama being president means that people voted for him, the second position is often predicative as in *A person may be a US president*, which should not lead us to conclude that *A person may be Barack Obama*.

For the inferences being made, we lack a temporal frame of reference. While this is less important for fictional individuals than for recognizable historical figures, it's necessary for understanding claims like *Possibly a US president accepts a nomination*, where the nomination happens before being elected president. Thus we rewrite the verbalizations to include both present and past tense readings: *Possibly a US president accepts (or accepted) a nomination*.

Cooper is a US president.
Possibly Cooper has (or had) an administration.

Emery is a scientist.
Possibly Emery finds (or has found) that something-is-the-case.

Williams is a writer.
Possibly Williams is (or was) successful.

Alex's car is a car.
It possibly has (or had) a seat.

Angel is a singer.
Possibly Angel is (or was) male.

Dolkhov is a dictator.
Possibly Dolkhov comes (or came) to power.

Gatanaia is a country.
Possibly it undergoes (or has undergone) an invasion.

Figure 6.1: **Selected examples from inference with unsharpened factoids.**

Beresford is a US president.
Probably Beresford raises (or raised) a tax.

Rene is a scientist.
Quite possibly Rene occasionally performs (or performed) a test.

Fournier is a writer.
Quite possibly Fournier occasionally undergoes (or underwent) translation.

Tremblay is a singer.
Quite possibly Tremblay occasionally performs (or performed) a song.

Patel's car is a car.
Probably it has (or had) a window as a part.

Lia is a river.
Quite possibly it occasionally overflows (or overflowed) a bank.

Avery is a dictator.
Quite possibly Avery occasionally invades (or invaded) a country.

Figure 6.2: **Selected examples from inference with sharpened factoids.**

	Judge 1	*Judge 2*	*Correlation*
Unsharp Q1	2.11	1.96	0.58
Unsharp Q2	1.94	2.16	0.64
Sharp Q1	1.81	1.77	0.57
Sharp Q2	2.50	2.68	0.49

Table 6.3: **Average ratings of inferences from factoids and sharpened axioms.** Q1 rates reasonableness from 1–5 with 1 being best. Q2 rates strength from (1) too weak to (5) too strong.

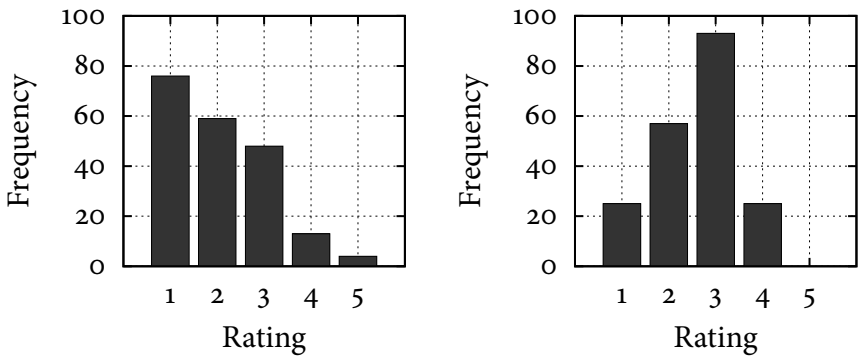

Figure 6.3: **Frequency of strength ratings for inferences with factoids and sharpened knowledge.** 3 is best.

For each of the ten classes of individuals used in the evaluation, the ten most frequently learned rules about that class were used for simple, one-step inferences. Then for each pair of a premise (*e.g.*, 'Emery is a scientist') and a conclusion drawn about the (gender-ambiguous) individual, two judges were asked first to rate whether the inference was reasonable and non-vacuous on a scale of 1–5, with 1 being full agreement and 5 being full disagreement. They then rated whether the strength of the conclusion was (1) much too weak (2) somewhat too weak (3) about right (4) somewhat too strong or (5) much too strong. The average ratings given by each judge are in Table 5, and histograms of the strength ratings are in Figure 1.

While the assessment of the reasonableness of the inferences made with the unsharpened factoids is quite favorable, it improves for the inferences made from sharpened axioms. This confirms that the sharpening process, which only produces axioms when a semantic rule matches the factoid, acts as an additional quality filter on the knowledge extraction. Looking at the histograms for the second question, rating whether the strength of the factoid is appropriate, we see a move toward 'about right' from 'much too weak'. However, there is still a general tendency to strengthen less than would be judged appropriate. For examples of inferences from both classes of knowledge, see Tables 3 and 4.

6.3.2 *Evaluating an Alternative Baseline*

While performing rough-and-ready inference with unsharpened factoids gives a weak baseline for evaluating the contribution of sharpening, we're also interested in how the entire knowledge extraction pipeline described compares with what is learned by other methods.

A prominent line of research in large-scale knowledge extraction looks for phrases identifying relations in text, using simple part-of-speech tagging and noun chunking rather than full syntactic parsing, and relying on the variety of ways that knowledge is expressed on the Web to find instances. The most notable tool in this vein is TextRunner (Banko *et al.*, 2007), which was compared with Knext by Van Durme & Schubert (2008).

More recently, Fader *et al.* (2011) introduced ReVerb, which attacks the problems of incoherent and uninformative extractions through the use of syntactic and lexical constraints. We use the publicly released ReVerb ClueWeb Extractions 1.1 data set, consisting of approximately 15 million binary assertions, as an alternative baseline.

As ReVerb is designed as an open *information* extraction system rather than an open *knowledge* extraction system like Lore, it discovers and stores many relations involving specific individuals, which are inappropriate as inferential knowledge. We prefiltered the ReVerb knowledge base to named entities and other, unknown lexical items. Then, as with the selection of factoids for inference, we chose the most frequent ReVerb formulas where the first position (arg1) is (a loose match of) the individual's class predicate. For a small comparison set, we used five of the predicate classes used for evaluating Lore and apply the ten most frequent rules per class.[3]

The resulting inferences were not judged on their strength but merely on how reasonable they were. Selected examples of the inferences are shown in Figure 6.4. The average ratings of inferences performed with ReVerb rules are 3.16 and 3.6 – significantly lower then for the inferences made with unsharpened factoids – with a 0.79 Pearson correlation between judges. While there is good knowledge, there are many extractions in the collection that apply only to a specific individual or scenario rather than more generally to the class, making for unlikely and confusing inferences.

6.3.3 *Discussion*

Our evaluation indicates that implicit knowledge learned from text, abstracted from statements about individuals, provides a more appropriate basis for simple, commonsense inferences than does the explicitly stated relations found by open information extraction systems like ReVerb. Furthermore, using lexico-syntactic rules to sharpen this

3 While ReVerb associates a confidence value with its extractions, the released collection all have confidence > 0.9. It wasn't a strong indicator of quality, but nor was frequency, which was noticeably skewed by copied spam text and boilerplate language from the Web corpus.

Rory is a US president.
Rory is the head of the US Armed Forces.

Rory is a US president.
Rory has visited Brazil.

Jessie is a scientist.
Jessie is working on a vaccine.

Devyn's car is a car.
It runs on water.

Smith is an artist.
Smith must be at least 18 years of age.

Griffiths is a dictator.
Griffiths cannot be good for English cricket.

Griffiths is a dictator.
Griffiths is a ruler with complete control.

Figure 6.4: **Selected examples of inferences with ReVerb extractions.**

knowledge into full logical forms allows for systematic improvements in the reasonableness of the conclusions drawn as well as the strength with which they are asserted.

It is important to note that we do not claim that this method is appropriate for learning all manner of useful commonsense knowledge. Much commonsense knowledge, such as script information or the usual outcomes of actions is best found from larger, intersentential patterns. Other knowledge, such as hyponymy is quite efficiently found through the application of simple patterns to large sources of text such as the Web.

The importance of the method described is in getting at implicit knowledge that is not readily found in other resources or by other techniques. And while we've evaluated only simple inferences here, where we see usefulness in this work is in the eventual use of commonsense knowledge in chaining and in combination with specific, domain knowledge.

6.4 Chapter Summary

We've described a pipeline for acquiring knowledge from patterns of predication in text and sharpening it to a form that is suitable for reasoning. The types of rules we obtain show a great deal of variety in the types of predicates they involve, as a result of the fact that even a single sentence can (and often does) yield one or more general quantified rules. Reasoning with sharpened knowledge rates favorably in human judgements of reasonableness and appropriate strength when compared with baseline inferences using unsharpened factoids or binary relations learned by ReVerb. Furthermore, we have shown at least in a preliminary way that the rules obtained are usable for inference and can provide probabilistically qualified conclusions.

7 Conclusions

> 'You know, when I was a lad they called it AI. Artificial Intelligence.'
>
> Hackworth allowed himself a tight, narrow, and brief smile. 'Well, there's something to be said for cheekiness, I suppose.'
>
> NEAL STEPHENSON, *The Diamond Age*, 1995

7.1 Summary

In Artificial Intelligence, it seems that the human-like understanding and reasoning required for problems such as question-answering, recognizing textual entailment, and planning depends on access to a large amount of knowledge. We have considered the representational requirements of that knowledge and the developments in logic, semantics, and AI that lead to us to the use of Episodic Logic, an expressive, natural language–like formalism.

Work in knowledge acquisition has ranged from the manual work of knowledge engineers to fully automated tools running over Web-scale text. For the learning of commonsense knowledge, not limited to predefined relations or individuals, we find that with filtering even the 'noisy' text that can be found on the Web is a good target for extraction. However, even over large volumes of text, much of the commonsense knowledge we seek is rarely expressed, at least as the explicit content of sentences. This problem of reporting bias motivates a focus on implicit knowledge and patterns of 'disconfirmed expectations' that let us learn what people presume to be true or expect to happen.

This approach to knowledge extraction is implemented in Lore, a tool that finds initial logical forms expressing what is possible in the world and then uses lexical-semantic rules and corpus frequencies to sharpen this knowledge into appropriately strong, partially disambiguated formulas that can be used for inference. The resulting knowledge uses generalized quantifiers like *most people*, which allow us to draw uncertain conclusions of varying strengths. The resulting inferences are rated better than a baseline performing pseudo-inference with unsharpened factoids or the results of a state-of-the-art information extraction system.

The appendices that follow present two lines of work that are alluded to in this dissertation: the use of crowdsourcing for the evaluation of knowledge bases and the interpretation of WordNet to acquire lexical axioms that complement what is learned from text.

7.2 Knowledge for Text Understanding

In this dissertation I have argued that better text understanding yields better collections of knowledge, and I trust that better knowledge will, in turn, enable better text understanding. When we understand a text, we make *bridging inferences* to connect consecutive sentences. Or, to frame the process more generally, we connect the contents of each sentence we read with what we already know. As Clark & Harrison (2010) put it, 'Prior knowledge should guide interpretation of new text, and new interpretations should augment that prior knowledge.' These inferences depend both on what has already been read and understood in that text and on our store of commonsense knowledge. For instance, Schank (1975) gives the sinister example:

> John wanted to become chief supervisor at the plant. He decided to go and get some arsenic.

People have no difficulty recognizing the intention that underlies the second sentence given the first, but when we try to enable a machine to do so, we find it requires

considerable knowledge. As discussed in §1.2, Schank suggests the core of this knowledge should be in the form of standard *scripts* and generic *plans*. But, regardless of representation, it seems we need to know, at the least, that:

1 'Chief supervisor' is a kind of position.
2 If a person dies, any position he or she holds becomes vacant.
3 A vacant position can be filled.
4 Arsenic is a poison.
5 If a person consumes poison, he or she may die.
6 If a person 'goes to get' something, he or she comes into possession of it.
7 If a person is in possession of something, he or she can use it.
8 Feeding someone something is using it.
9 If a person wants something, he or she will often take actions to make it happen.

While some of these are attainable by the methods presented in this dissertation (*e.g.*, *Arsenic is a poison* and *If a person consumes a poison, he or she may die*), others are so basic they are unlikely to be learned in this way.

Even without goals of 'deep' natural language understanding, simply resolving pronouns requires chains of reasoning. For instance, Minsky (1974) gives an example of an elementary school story about Jane considering buying a kite for Jack. Her friend Penny tells her, 'He already has a Kite. He will make you take it back.' Here, 'it' doesn't refer to the most recently mentioned thing (the kite Jack already has) but to the new kite Jane might buy him. Similarly, Lenat (1995) contrasted the referent of 'they' in 'The police arrested the demonstrators because they feared violence' *vs* 'The police arrested the demonstrators because they advocated violence'. It is our knowledge of how police behave that guides our interpretation of the sentence. As Singh (2002) observes, 'people seem to need a tremendous amount of knowledge of a very diverse variety to understand even the simplest children's story.'

7.3 Future Work

> We can only see a short distance ahead, but we can see plenty there that needs to be done.
>
> ALAN TURING, 'Computing Machinery and Intelligence', 1950

In this section, I briefly describe some steps toward better commonsense knowledge to support text understanding and AI.

Abstracting knowledge to appropriate generality Is a claim specific to an individual or a generic claim about a class? How broad should the class be? Knowledge base abstraction helps ensure that we store knowledge at the level where it is most interesting for inference. For instance, if I tell you that John is a painter, there's no cognitive or practical reason to think of his aorta. Additionally, idiosyncratic properties of a single – possibly frequently mentioned – instance should not be allowed to dominate what we know about a class. *E.g.*, it's fine to learn that *Bill Clinton may occasionally play a saxophone*, but it is unhelpful to abstract this to a claim about US presidents generally. To this end, we can compute the mutual information between the predication and 'Bill Clinton' *vs* 'US president'.

Identifying alternatives It is possible for Lore to learn contradictory knowledge, *e.g.*, *Most house cats are black* and *Most house cats are calicos*. Contradictory claims can be minimized by recognizing which nominal and adjectival predicates are mutually exclusive alternatives. However, as Minsky (1974) wrote, 'the preoccupation with Consistency, so valuable for Mathematical Logic, has been incredibly destructive to those working on models of mind.' The possibility of deriving contradictory conclusions with some uncertainty should not dissuade us from learning what we can from text.

Learning from intersentential discourse While my work has focused on knowledge implicit in individual sentences, there is an abundance of knowledge – especially about

the results of actions – that can be found in sequential sentences, *e.g.*, 'I dropped the plate. There were shards everywhere.' Contrast, however, 'I dropped the plate. The doorbell rang', where the implicit intersentential relationship is temporal, not causal. A related issue is the introduction of coreference/anaphora resolution, which has the potential to lead to more specific claims – rather than abstract 'she' to *a female*, we might be able to abstract to *a singer*, *an empress*, *etc.*

Interpretation of generic sentences Commonsense knowledge can be learned by abstracting from direct experience with the world or, in this thesis, the world of text. It can also be learned from the explicit statement of generalizations. These can be found in intentionally informative sources like encyclopedias or in the data of the Open Mind project, which solicited statements like 'Books are used to learn things'. This knowledge tends to be expressed as *generic sentences*, which, as Krifka *et al.* (1995) describe, abstract away from particular objects to genera or kinds and from particular events and facts to regularities about groups of episodes, events, or states of affairs. While a generic sentence can superficially be identical to a particular sentence (*e.g.*, 'Lions stalk their prey'), Reiter & Frank (2010) present promising work on identifying generics.

As in sharpening factoids, a key issue when forming axioms from generic sentences is radical underspecification: What, for instance, do you do with a book in order to learn? When we say that 'A bird lays eggs', what is the domain restrictor? It is at most *female* birds. Many generic sentences specify a relation between an antecedent situation and a consequent situation. *E.g.*, we interpret 'Hurrying causes accidents' to mean that in the situations in which one is performing some task, an accident will occur more often in those where one is hurrying than in those where one is not.

This situational analysis becomes more complicated if we consider the meaning of a statement like 'Smoking kills people', where the consequent is not required to happen in immediate temporal connection with the antecedent: One can smoke for years, stop smoking, live twenty years, and then die of side-effects of smoking. Here, the antecedent is a habitual. We also find causal sentences like 'kittens cause happiness'. This could be

understood as underspecification meaning something like 'being in a situation where there is a kitten causes happiness'. The same analysis could be made for 'Joe caused the fire': We understand the sentence to mean that Joe is the agent of some particular action that caused the fire, *e.g.*, 'Joe kicking over the lamp caused the fire'.

Enumerating fundamental abstract knowledge Improving performance at these problems will also highlight knowledge less readily discovered from text, including some of the most basic psychological, causal, and spatiotemporal rules, such as those listed in §7.2. Manually engineering these is a practical alternative to building Turing's 'child machine' – a robot with human-like perceptual and motor abilities that could learn these basic notions in a 'situated' way through interaction with the world.

7.4 Final Remarks

The work described in this dissertation is an effort to create the knowledge infrastructure needed for natural language understanding and commonsense reasoning. By looking for the knowledge and expectations presumed in textual discourse and by applying semantic criteria to large collections of shallow knowledge, I produce collections of high-quality knowledge, appropriate for reasoning. Yet there is much more a machine needs to know to enable human-like natural language understanding and commonsense reasoning, motivating future work on learning by reading.

References

A. Akbik and A. Löser. Kraken: N-ary facts in open information extraction. In J. Fan, R. Hoffman, A. Kalyanpur, S. Riedel, F. M. Suchanek, and P. P. Talukdar, editors, *Proceedings of the NAACL Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction* (*AKBC-WEKEX*), pages 52–6, Montréal, Quebec, Canada, June 2012. Association for Computational Linguistics.

J. F. Allen. *Natural Language Understanding*. Benjamin/Cummings, Redwood City, CA, 2nd edition, 1994.

J. Álvez, J. Atserias, J. Carrera, S. Climent, E. Laparra, A. Oliver, and G. Rigau. Complete and consistent annotation of WordNet using the Top Concept Ontology. In *Proceedings of the Language Resource and Evaluation Conference* (*LREC*), pages 1529–34, 2008.

J. Álvez, P. Lucio, and G. Rigau. Adimen-SUMO: Reengineering an ontology for first-order reasoning. *International Journal on Semantic Web and Information Systems*, 8 (4), 2012.

R. A. Amsler and J. S. White. Development of a computational methodology for deriving natural language semantic structures via analysis of machine-readable dictionaries. Technical Report MCS 77-01315, National Science Foundation, 1979.

S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (*LREC*), 2010.

T. Baldwin and F. Bond. A plethora of methods for learning English countability. In M. Collins and M. Steedman, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 73–80, Sapporo, Japan, July 2003. Association for Computational Linguistics.

M. Banko and O. Etzioni. Strategies for lifelong knowledge extraction from the Web. In D. H. Sleeman and K. Barker, editors, *Proceedings of the Fourth International Conference on Knowledge Capture* (*K-CAP*), pages 95–102, Whistler, British Columbia, Canada, Oct. 2007. Association for Computing Machinery.

M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the Web. In M. M. Veloso, editor, *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence* (IJCAI), pages 2670–6, Hyderabad, India, 2007.

R. Bar-Haim and I. Dagan. Efficient semantic inference over language expressions. Talk at NSF Symposium on Semantic Knowledge Discovery, Organization, and Use, 2008.

K. Barker, B. Porter, and P. Clark. A library of generic concepts for composing knowledge bases. In *Proceedings of the First International Conference on Knowledge Capture* (K-CAP), pages 14–21, Victoria, British Columbia, Canada, Oct. 2001. Association for Computing Machinery.

J. Barwise and J. Perry. *Situations and Attitudes*. MIT Press, Cambridge, MA, 1983.

R. Barzilay and L. Lee. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In M. Hearst and M. Ostendorf, editors, *Human Language Technologies: Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics* (HLT-NAACL), Edmonton, Alberta, Canada, May–June 2003.

BNC Consortium. The British National Corpus, v. 2. Distributed by Oxford University Computing Services, 2001. URL www.natcorp.ox.ac.uk.

F. Bond and C. Vatikiotis-Bateson. Using an ontology to determine English countability. In *Proceedings of the International Conference on Computational Linguistics* (COLING), pages 1–7, 2002.

J. L. Borges. Funes the memorious. *La Nación*, June 1942. Reprinted in English translation in *Labyrinths*, 1962.

J. Boyd-Graber, C. Fellbaum, D. Osherson, and R. Schapire. Adding dense, weighted connections to WordNet. In *Proceedings of the Third Global WordNet Meeting*, 2006.

T. Brants and A. Franz. Web 1T 5-gram version 1. Distributed by the Linguistic Data Consortium, 2006.

H. C. Bunt. *Mass Terms and Model-Theoretic Semantics*. Cambridge Studies in Linguistics Series. Cambridge University Press, 1985.

K. Burton, A. Java, and I. Soboroff. The ICWSM 2009 Spinn3r dataset. In E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, editors, *Proceedings of the Third International Conference on Weblogs and Social Media* (ICWSM), San Jose, CA, May 2009.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an architecture for never-ending language learning. In M. Fox and D. Poole, editors, *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence* (AAAI), pages 1306–13, Atlanta, GA, July 2010. AAAI Press.

G. N. Carlson. *Reference to Kinds in English*. Ph.D. in linguistics, University of Massachussetts, Amherst, 1977a. Published 1980, Garland Press, New York.

———. A unified analysis of the English bare plural. *Linguistics and Philosophy*, 1(3): 413–57, 1977b.

———. Generic terms and generic sentences. *Journal of Philosophical Logic*, 11:145–81, 1982.

———. Truth-conditions of generic sentences: Two contrasting views. In G. N. Carlson and F. J. Pelletier, editors, *The Generic Book*, Studies in Communication, Media, and Public Opinion, pages 224–37. University of Chicago Press, Chicago, IL, 1995.

L. Carroll. *Through the Looking-Glass, and What Alice Found There*. Macmillan, London, 1871. Illustrated by John Tenniel.

N. Chambers and D. Jurafsky. Unsupervised learning of narrative event chains. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics* (ACL): *Human Language Technologies* (HLT), pages 789–97, Columbus, OH, June 2008.

———. Unsupervised learning of narrative schemas and their participants. In K. Y. Su, J. Su, and J. Wiebe, editors, *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics* (ACL), pages 602–10, Suntec, Singapore, Aug. 2009. Association for Computational Linguistics.

E. Charniak. A maximum-entropy-inspired parser. In J. Wiebe, editor, *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics* (NAACL), pages 132–9, Seattle, WA, Apr.–May 2000. Morgan Kaufmann Publishers.

G. Chierchia and S. McConnell-Ginet. *Meaning and Grammar: An Introduction to Semantics*. MIT Press, Cambridge, MA, 2nd edition, 2000.

T. Chklovski. *Using Analogy to Acquire Commonsense Knowledge from Human Contributors*. PhD thesis, MIT Artificial Intelligence Laboratory, Feb. 2003.

T. Chklovski and P. Pantel. VerbOcean: Mining the Web for fine-grained semantic verb relations. In D. Lin and D. Wu, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 33–40, Barcelona, Spain, July 2004. Association for Computational Linguistics.

M. S. Chodorow, R. J. Byrd, and G. E. Heidorn. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (ACL), pages 299–304, University of Chicago, Chicago, IL, July 1985.

A. Christie. *The Moving Finger*. Dodd, Mead and Company, New York, NY, July 1942.

H. H. Clark. Bridging. In R. C. Schank and B. L. Nash-Webber, editors, *Proceedings of the Workshop on Theoretical Issues in Natural Language Processing*, pages 169–74, Cambridge, MA, 1975. Association for Computational Linguistics.

P. Clark and P. Harrison. Large-scale extraction and use of knowledge from text. In Y. Gil and N. F. Noy, editors, *Proceedings of the Fifth International Conference on Knowledge Capture* (K-CAP), pages 153–60, Redondo Beach, CA, Sept. 2009a. Association for Computing Machinery.

———. An inference-based approach to recognizing entailment. In *Proceedings of the Text Analysis Conference* (TAC), Gaithersburg, MD, Nov. 2009b.

———. Machine reading as a process of partial question-answering. In R. Mulkar-Mehta, J. Allen, J. R. Hobbs, E. Hovy, B. Magnini, and C. Manning, editors, *Proceedings of the NAACL-HLT First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 1–9, Los Angeles, CA, June 2010. Association for Computational Linguistics.

P. Clark, C. Fellbaum, and J. R. Hobbs. Using and extending WordNet to support question-answering. In A. Tanács, D. Csendes, V. Vincze, C. Fellbaum, and P. Vossen, editors, *Proceedings of the Fourth Global WordNet Conference* (GWC), pages 111–9, University of Szeged, Hungary, Jan. 2008a.

P. Clark, C. Fellbaum, J. R. Hobbs, P. Harrison, W. R. Murray, and J. Thompson. Augmenting WordNet for deep understanding of text. In J. Bos and R. Delmonte, editors, *Proceedings of the Symposium on Semantics in Text Processing* (STEP), volume 1 of *Research in Computational Semantics*, pages 45–57, Venice, Italy, 2008b. College Publications.

D. R. Clausen and C. D. Manning. Presupposed content and entailments in natural language inference. In *Proceedings of the ACL-IJCNLP Workshop on Applied Textual Inference*, pages 70–3, 2009.

A. Cohen. Generics, frequency adverbs, and probability. *Linguistics and Philosophy*, 22 (3):221–53, 1999.

M. Collins. Three generative, lexicalised models for statistical parsing. In P. R. Cohen and W. Wahlster, editors, *Proceedings of the 35th Annual Meeting of the Association*

for Computational Linguistics (ACL), pages 16–23, Madrid, Spain, July 1997. Morgan Kaufmann Publishers / Association for Computational Linguistics.

J. Cowie and W. Lehnert. Information extraction. *Communications of the Association for Computing Machinery*, 39:80–91, Jan. 1996.

J. Curtis, D. Baxter, P. Wagner, J. Cabral, D. Schneider, and M. J. Witbrock. Methods of rule acquisition in the TextLearner system. In S. Nirenburg and T. Oates, editors, *Proceedings of the AAAI Spring Symposium on Learning by Reading and Learning to Read*, pages 22–8. AAAI Press, Mar. 2009.

D. Davidson. The logical form of action sentences. In N. Rescher, editor, *The Logic of Decision and Action*. University of Pittsburgh Press, 1967. Reprinted in D. Davidson and G. Harman, editors, *The Logic of Grammar*, pages 235–45, Dickenson Publ., Encino, CA, 1975.

US Department of Transportation. National transportation statistics, Oct. 2009. URL www.bts.gov/publications/national_transportation_statistics.

J. R. Doppa, M. NasrEsfahani, M. S. Sorower, T. G. Dietterich, X. Fern, and P. Tadepalli. Towards learning rules from natural texts. In R. Mulkar-Mehta, J. Allen, J. R. Hobbs, E. Hovy, B. Magnini, and C. Manning, editors, *Proceedings of the NAACL-HLT First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 70–7, Los Angeles, CA, June 2010. Association for Computational Linguistics.

D. Downey, O. Etzioni, and S. Soderland. A probabilistic model of redundancy in information extraction. In L. P. Kaelbling and A. Saffiotti, editors, *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence* (IJCAI), pages 1034–41, Edinburgh, Scotland, July–August 2005. Professional Book Center.

D. Downey, O. Etizioni, and S. Soderland. Analysis of a probabilistic model of redundancy in unsupervised information extraction. *Artificial Intelligence*, 174(11):726–48, July 2010.

O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Web-scale information extraction in KnowItAll. In *Proceedings of the Thirteenth International World Wide Web Conference*, 2004.

A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In P. Merlo, R. Barzilay, and M. Johnson, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

C. J. Fillmore and C. Baker. A frames approach to semantic analysis. In B. Heine and H. Narrog, editors, *The Oxford Handbook of Linguistic Analysis*. 2009.

K. Forbus, C. Riesbeck, L. Birnbaum, K. Livingston, A. Sharma, and L. Ureel. Integrating natural language, knowledge representation and reasoning, and analogical processing to learn by reading. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence* (*AAAI*), pages 1542–7, Vancouver, British Columbia, Canada, July 2007. AAAI Press.

K. Forbus, K. Lockwood, A. Sharma, and E. Tomai. Steps toward a 2nd generation learning by reading system. In S. Nirenburg and T. Oates, editors, *Proceedings of the AAAI Spring Symposium on Learning by Reading and Learning to Read*, pages 36–43. AAAI Press, Mar. 2009.

N. S. Friedland and P. G. Allen. The Halo pilot: Towards a digital Aristotle. Technical report, Vulcan, Inc., 2003. URL projecthalo.com/content/docs/halopilot_vulcan_finalreport.pdf.

A. Gangemi, R. Navigli, and P. Velardi. The OntoWordNet project: Extension and axiomatization of conceptual relations in WordNet. In *Proceedings of Ontologies, Databases, and Applications of Semantics* (*ODBASE*), pages 820–38, Catania, Sicily, Italy, 2003. Springer-Verlag.

R. Girju. Automatic detection of causal relations for question answering. In *Proceedings of the ACL Workshop on Multilingual Summarization and Question Answering*, pages 76–83, Stroudsburg, PA, 2003. Association for Computational Linguistics.

A. Gordon and R. Swanson. Identifying personal stories in millions of weblog entries. In *Proceedings of the Third International Conference on Weblogs and Social Media* (*ICWSM*), *Data Challenge Workshop*, May 2009.

J. M. Gordon and L. K. Schubert. Quantificational sharpening of commonsense knowledge. In C. Havasi, D. B. Lenat, and B. Van Durme, editors, *Proceedings of the AAAI Fall Symposium on Commonsense Knowledge* (*CSK*), Arlington, VA, Nov. 2010. AAAI Press.

———. Discovering commonsense entailment rules implicit in sentences. In S. Pado and S. Thater, editors, *Proceedings of the EMNLP Workshop on Textual Entailment* (*TextInfer*), Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

———. Using textual patterns to learn expected event frequencies. In J. Fan, R. Hoffman, A. Kalyanpur, S. Riedel, F. M. Suchanek, and P. P. Talukdar, editors, *Proceedings of the NAACL Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction* (*AKBC-WEKEX*), pages 122–7, Montréal, Quebec, Canada, June 2012. Association for Computational Linguistics.

———. WordNet hierarchy axiomatization and the mass–count distinction. In D. A. Evans, M. van der Schaar, and P. Sheu, editors, *Proceedings of the Seventh International Conference on Semantic Computing* (ICSC), Irvine, CA, Sept. 2013. IEEE.

J. M. Gordon and B. Van Durme. Reporting bias and knowledge acquisition. In F. M. Suchanek, S. Riedel, S. Singh, and P. P. Talukdar, editors, *Proceedings of the Workshop on Automated Knowledge Base Construction* (AKBC), pages 23–30, San Francisco, CA, Oct. 2013. Association for Computing Machinery.

J. M. Gordon, B. Van Durme, and L. K. Schubert. Weblogs as a source for extracting general world knowledge. In Y. Gil and N. F. Noy, editors, *Proceedings of the Fifth International Conference on Knowledge Capture* (K-CAP), pages 185–6, Redondo Beach, CA, Sept. 2009. Association for Computing Machinery.

———. Evaluation of commonsense knowledge with Mechanical Turk. In C. Callison-Burch and M. Dredze, editors, *Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 159–62, Los Angeles, CA, June 2010a. Association for Computational Linguistics.

———. Learning from the Web: Extracting general world knowledge from noisy text. In V. Nastase, R. Navigli, and F. Wu, editors, *Proceedings of the AAAI Workshop on Collaboratively-built Knowledge Sources and Artificial Intelligence*, pages 8–13, Atlanta, GA, July 2010b. AAAI Press.

D. Graff, J. Kong, K. Chen, and K. Maeda. English Gigaword. Distributed by the Linguistic Data Consortium, 2007.

H. P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, San Diego, CA, 1975.

N. Guarino and C. Welty. Ontological analysis of taxonomic relationships. In A. Laender and V. Storey, editors, *Proceedings of the 19th International Conference on Conceptual Modeling*. Springer-Verlag, 2000.

A. Gusev, N. Chambers, P. Khaitan, D. Khilnani, S. Bethard, and D. Jurafsky. Using query patterns to learn the duration of events. In *Proceedings of the Ninth International Conference on Computational Semantics* (IWCS), pages 145–154, Oxford, England, 2011. Association for Computational Linguistics.

S. Hamm. Watson on Jeopardy!, Feb. 2011. URL asmarterplanet.com/blog/2011/02/watson-on-jeopardy-day-one-man-vs-machine-for-global-bragging-rights.html.

Z. Harris. Distributional structure. In J. J. Katz, editor, *The Philosophy of Linguistics*, pages 26–47. Oxford University Press, New York, NY, 1985.

M. S. Hart and volunteers. Project Gutenberg, 2006. URL www.gutenberg.org.

C. Havasi, R. Speer, and J. Alonso. ConceptNet 3: A flexible, multilingual semantic network for common sense knowledge. In R. Mitkov, editor, *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (RANLP), Borovets, Bulgaria, Sept. 2007.

M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics* (COLING), pages 539–45, 1992.

J. R. Hobbs. Ontological promiscuity. In *Proceedings of the Twenty-Third Annual Meeting of the Association for Computational Linguistics* (ACL), pages 61–9, Chicago, IL, 1985. Association for Computational Linguistics.

———. Discourse and inference, 2013. URL isi.edu/~hobbs/disinf-tc.html. Unpublished manuscript.

J. R. Hobbs and A. Gordon. Encoding knowledge of commonsense psychology. In *Proceedings of the Seventh International Symposium on Logical Formalizations of Commonsense Reasoning*, Corfu, Greece, May 2005.

R. Hoffman, S. Amershi, K. Patel, F. Wu, J. Fogarty, and D. S. Weld. Amplifying community content creation with mixed-initiative information extraction. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI), 2009.

C. H. Hwang and L. K. Schubert. EL: A formal, yet natural, comprehensive knowledge representation. In *Proceedings of the Eleventh National Conference on Artificial Intelligence* (AAAI-93), pages 676–82, Washington, DC, July 1993.

———. Interpreting tense, aspect, and time adverbials: a compositional, unified approach. In D. M. Gabbay and H. J. Ohlbach, editors, *Proceedings of the International Conference on Temporal Logic*, pages 238–64, Bonn, Germany, 1994. Springer-Verlag.

N. Ide and J. Véronis. Knowledge extraction from machine-readable dictionaries: An evaluation. In P. Steffens, editor, *Machine Translation and the Lexicon*. Springer-Verlag, Berlin, Germany, 1994.

R. Izquierdo, A. Suárez, and G. Rigau. Exploring the automatic selection of basic level concepts. In R. Mitkov, editor, *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (RANLP), Borovets, Bulgaria, Sept. 2007.

R. Jackendoff. Parts and boundaries. *Cognition*, 41:9–45, 1991.

A. N. Kaplan and L. K. Schubert. Measuring and improving the quality of world knowledge extracted from WordNet. Technical report, University of Rochester, Rochester, NY, 2001.

L. Karttunen. Presuppositions of compound sentences. *Linguistic Inquiry*, 4:169–93, 1973.

P. Kingsbury and M. Palmer. PropBank: The next level of TreeBank. In *Proceedings of Treebanks and Lexical Theories 2003*, 2003.

K. Kipper-Schuler. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2006.

T. Kiss, F. J. Pelletier, and T. Stadtfeld. Building a reference lexicon for countability in English. In *Proceedings of the Language Resources and Evaluation Conference* (LREC), Reykjavik, Iceland, May 2014.

A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, pages 453–6, Florence, Italy, 2008.

M. Krifka, F. J. Pelletier, G. N. Carlson, A. ter Meulen, G. Chierchia, and G. Link. Genericity: An introduction. In G. N. Carlson and F. J. Pelletier, editors, *The Generic Book*, Studies in Communication, Media, and Public Opinion, pages 1–124. University of Chicago Press, Chicago, IL, 1995.

H. Kučera and W. N. Francis. *Computational Analysis of Present-Day American English*. Brown University Press, 1967.

M. Lapata and F. Keller. The Web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proceedings of Human Language Technologies – North American Chapter of the Association for Computational Linguistics* (HLT-NAACL), pages 121–8, 2004.

———. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2(1):1–31, 2005.

D. B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the Association for Computing Machinery*, 38(11):33–48, 1995.

C. I. Lewis. *A Survey of Symbolic Logic*. University of California Press, Berkeley, CA, 1918.

C. S. Lewis. *The Voyage of the Dawn Treader*. Geoffrey Bles, London, 1952.

D. Lin and P. Pantel. DIRT: Discovery of inference rules from text. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 323–8, 2001.

———. Concept discovery from text. In *Proceedings of the 19th International Conference on Computational Linguistics* (COLING), pages 1–7, Taipei, Taiwan, Aug.–Sept. 2002. Association for Computational Linguistics.

H. Liu and P. Singh. Commonsense reasoning in and over natural language. In *Proceedings of the Eighth International Conference on Knowledge-Based Intelligent Information and Engineering Systems* (KES). Springer-Verlag, Sept. 2004.

B. MacCartney and C. D. Manning. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics* (COLING), Manchester, England, 2008.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–30, 1994.

W. Mason and D. J. Watts. Financial incentives and the 'performance of crowds'. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 77–85, Paris, France, 2009. Association for Computing Machinery.

C. Matuszek, M. Witbrock, R. C. Kahlert, J. Cabral, D. Schneider, P. Shah, and D. Lenat. Searching for common sense: Populating Cyc from the Web. In *Proceedings of the Twentieth National Conference on Artificial Intelligence*, 2005.

C. Matuszek, J. Cabral, M. Wittbrock, and J. DeOliveira. An introduction to the syntax and content of Cyc. In *Proceedings of the AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, Stanford, CA, Mar. 2006.

J. McCarthy. Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, pages 75–91, London, England, 1959. Her Majesty's Stationary Office.

———. Circumscription – a new form of non-monotonic reasoning. *Artificial Intelligence*, 13(1–2), 1980. Reprinted in B.L. Webber and N.J. Ginsberg (ed.), *Readings in Nonmonotonic Reasoning*, Morgan Kaufmann, 1987, pages 145–52.

———. Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence*, 28:89–116, 1986.

———. Artificial intelligence, logic, and formalizing common sense. In *Philosophical Logic and Artificial Intelligence*, pages 161–90. Kluwer Academic Publishers, 1990.

J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969.

D. McClosky, E. Charniak, and M. Johnson. Effective self-training for parsing. In *Proceedings of the Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics* (HLT-NAACL), New York, NY, 2006.

G. A. Miller and F. Hristea. WordNet nouns: Classes and instances. *Computational Linguistics*, 32(1):1–3, 2006.

M. Minsky. A framework for representing knowledge. Technical Report AI Lab Memo 306, Massachussetts Institute of Technology, Cambridge, MA, June 1974. Reprinted in abridged form in P.H. Winston, ed., *The Psychology of Computer Vision*, pages 211–77, McGraw-Hill, 1975.

S. Moody. The brain behind Cyc. *The Austin Chronicle*, 1999. http://www.austinchronicle.com/screens/1999-12-24/75252.

F. Morbini and L. K. Schubert. Evaluation of Epilog: A reasoner for Episodic Logic. In *Proceedings of the International Symposium on Logical Formalizations of Commonsense Reasoning*, pages 103–8, Toronto, Canada, June 2009.

C. Napoles, M. Gormley, and B. Van Durme. Annotated Gigaword. In J. Fan, R. Hoffman, A. Kalyanpur, S. Riedel, F. M. Suchanek, and P. P. Talukdar, editors, *Proceedings of the NAACL Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction* (AKBC-WEKEX), pages 95–100, Montréal, Quebec, Canada, June 2012. Association for Computational Linguistics.

D. Nicolas. The logic of mass expressions. In E. N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. Center for the Study of Language and Information, Stanford University, Stanford, CA, 2014. URL plato.stanford.edu/entries/logic-massexpress.

I. Niles and A. Pease. Toward a standard upper ontology. In C. Welty and B. Smith, editors, *Proceedings of the International Conference on Formal Ontology in Information Systems* (FOIS), 2001.

T. O'Hara, N. Salay, M. Witbrock, D. Schneider, B. Aldag, S. Bertolo, K. Panton, F. Lehmann, M. Smith, D. Baxter, J. Curtis, and P. Wagner. Inducing criteria for mass noun lexical mappings using the Cyc KB, and its extension to WordNet. In *Proceedings of the International Workshop on Computational Semantics* (IWCS), pages 34–9, Apr.–May 2003.

P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (ACL), 2006.

P. Pantel, R. Bhagat, B. Coppola, T. Chklovski, and E. H. Hovy. ISP: Learning inferential selectional preferences. In C. L. Sidner, T. Schultz, M. Stone, and C. Zhai, editors, *Human Language Technologies: Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), Rochester, NY, Apr. 2007. Association for Computational Linguistics.

T. Parsons. *Events in the Semantics of English*. MIT Press, Cambridge, MA, 1990.

J. Pearl. Probabilistic semantics for nonmonotonic reasoning. In R. Cummins and J. Pollock, editors, *Philosophy and AI: Essays at the Interface*. Bradford Books, 1995.

F. J. Pelletier. Non-singular reference: Some preliminaries. *Philosophia*, 5:451–65, 1975. Reprinted in Pelletier (1979).

F. J. Pelletier, editor. *Mass Terms: Some Philosophical Problems*. Kluwer Academic Publishers, Dordrecht, the Netherlands, 1979.

F. J. Pelletier and L. K. Schubert. Mass expressions. In *Handbook of Philosophical Logic*, volume 10, pages 327–407. Reidel, Dordrecht, the Netherlands, 1989. Expanded version reprinted in second edition (2003), volume 10, pp. 265–350.

A. Purtee and L. K. Schubert. TTT: A tree transduction language for syntactic and semantic processing. In *Proceedings of the EACL Workshop on Applications of Tree Automata Techniques in Natural Language Processing*, pages 21–30, Avignon, France, Apr. 2012.

J. Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, 1995.

J. Pustejovsky, C. Havasi, J. Littman, A. Rumshisky, and M. Verhagen. Towards a generative lexical resource: The Brandeis Semantic Ontology. In *Proceedings of the Language Resource and Evaluation Conference* (*LREC*), Genoa, Italy, 2006.

S. Raghavan and R. J. Mooney. Online inference-rule learning from natural-language extractions. In *Proceedings of the AAAI Workshop on Statistical Relational AI* (*StaRAI-13*), July 2013.

H. Reichenbach. *Elements of Symbolic Logic*. Macmillan, New York, NY, 1947.

N. Reiter and A. Frank. Identifying generic noun phrases. In J. Hajic, S. Carberry, and S. Clark, editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 40–9, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

M. Roemmele, C. A. Bejan, and A. S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford University, Mar. 2011.

D. Rohde. TGrep2 manual, 2001. Unpublished manuscript, Brain & Cognitive Science Department, MIT.

S. Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1:261–377, Mar. 2008.

S. A. Schaeffer, C. H. Hwang, J. de Haan, and L. K. Schubert. Epilog, the computational system for Episodic Logic: User's guide. Technical report, Department of Computing Science, University of Alberta, Aug. 1993.

R. C. Schank. Using knowledge to understand. In R. C. Schank and B. L. Nash-Webber, editors, *Proceedings of the Workshop on Theoretical Issues in Natural Language Processing*, pages 117–21, Cambridge, MA, 1975. Association for Computational Linguistics.

S. Schoenmackers, O. Etzioni, D. S. Weld, and J. Davis. Learning first-order Horn clauses from Web text. In H. Li and L. Màrquez, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), East Stroudsburg, PA, Oct. 2010. Association for Computational Linguistics.

L. K. Schubert. Can we derive general world knowledge from texts? In *Proceedings of the Second International Conference on Human Language Technology Research* (*HLT*), pages 94–7, San Diego, CA, Mar. 2002. Morgan Kaufmann Publishers.

———. From generic sentences to scripts. In *Proceedings of the IJCAI Workshop on Logic and the Simulation of Interaction and Reasoning*, Pasadena, CA, 2009.

———. Computational linguistics. In E. N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. Center for the Study of Language and Information, Stanford University, Stanford, CA, 2014. URL plato.stanford.edu/entries/computational-linguistics.

L. K. Schubert and C. H. Hwang. An episodic knowledge representation for narrative texts. Technical Report 345, University of Rochester, 1990.

———. Episodic Logic meets Little Red Riding Hood: A comprehensive, natural representation for language understanding. In L. Iwanska and S. C. Shapiro, editors, *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. MIT/AAAI Press, 2000.

L. K. Schubert and F. J. Pelletier. From English to logic: Context free computation of 'conventional' logical translations. *American Journal of Computational Linguistics*, 8:26–44, 1982. Also in *Readings in Natural Language Processing*, B. Grosz, K. Jones, and B. Webber, eds., 293–311, Morgan Kaufman, Los Altos, CA, 1986.

L. K. Schubert and M. H. Tong. Extracting and evaluating general world knowledge from the Brown corpus. In *Proceedings of the HLT-NAACL Workshop on Text Meaning*, pages 7–13, Edmonton, Alberta, Canada, May 2003.

L. K. Schubert, B. Van Durme, and M. Bazrafshan. Entailment inference in a natural logic–like general reasoner. In C. Havasi, D. B. Lenat, and B. Van Durme, editors, *Proceedings of the AAAI Fall Symposium on Commonsense Knowledge* (*CSK*), Arlington, VA, Nov. 2010. AAAI Press.

S. Sekine. Automatic paraphrase discovery based on context and keywords between NE pairs. In *Proceedings of the Third International Workshop on Paraphrasing* (*IWP*), 2005.

S. Sekine, editor. *Notebook of the NSF Symposium on Semantic Knowledge Discovery, Organization, and Use*, New York University, Nov. 2008.

L. A. Seneca. On the shortness of life. In *Dialogues and Letters*. Penguin Classics, 1997. Translation by C.D.N. Costa.

B. Shannon. On the two kinds of presuppositions in natural language. *Foundations of Language*, 14:247–9, 1976.

J. Simpson, editor. *Oxford English Dictionary Online*. Oxford University Press, Oxford, Dec. 2013. URL www.oed.com.

P. Singh. The public acquisition of commonsense knowledge. In *Proceedings of the AAAI Spring Symposium on Acquiring* (*and Using*) *Linguistic* (*and World*) *Knowledge for Information Access*, Palo Alto, CA, 2002.

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast – but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 254–63, Honolulu, HI, Oct. 2008. Association for Computational Linguistics.

S. Sorower, T. G. Dietterich, J. R. Doppa, W. Orr, P. Tadepalli, and X. Fern. Inverting Grice's maxims to learn rules from natural language extractions. In *Proceedings of Neural Information Processing Systems* (*NIPS*), pages 1053–61, 2011.

P. Srinivasan and A. Yates. Quantifier scope disambiguation using extracted pragmatic knowledge: Preliminary results. In P. Koehn and R. Mihalcea, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 1465–74, Singapore, Aug. 2009. Association for Computational Linguistics.

N. Stephenson. *The Diamond Age: Or, A Young Lady's Illustrated Primer*. Bantam Spectra, New York, NY, 1995.

T. Stoppard. *Rosencrantz and Guildenstern are Dead*. 1966. First staged at the Edinburgh Festival Fringe. Published by Faber and Faber, London, 1967.

F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge unifying WordNet and Wikipedia. In *Proceedings of the Sixteenth International World Wide Web Conference* (*WWW*), pages 697–706, Banff, Canada, May 2007.

I. Szpektor, H. Tanev, I. Dagan, and B. Coppola. Scaling Web-based acquisition of entailment relations. In D. Lin and D. Wu, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Barcelona, Spain, July 2004. Association for Computational Linguistics.

I. Szpektor, E. Shnarch, and I. Dagan. Instance-based evaluation of entailment rule acquisition. In J. A. Carroll, A. van den Bosch, and A. Zaenan, editors, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 456–63, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

R. H. Thomason, editor. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven, CT, 1974.

A. M. Turing. Computing machinery and intelligence. *Mind*, LIX, Oct. 1950.

B. Van Durme. *Extracting Implicit Knowledge from Text*. PhD thesis, University of Rochester, 2010.

B. Van Durme and L. K. Schubert. Open knowledge extraction through compositional language processing. In J. Bos and R. Delmonte, editors, *Proceedings of the Symposium on Semantics in Text Processing* (*STEP*), volume 1 of *Research in Computational Semantics*, pages 239–54, Venice, Italy, 2008. College Publications.

B. Van Durme, T. Qian, and L. K. Schubert. Class-driven attribute extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics* (*COLING*), pages 921–8, Manchester, UK, Aug. 2008.

B. Van Durme, P. Michalak, and L. K. Schubert. Deriving generalized knowledge from corpora using WordNet abstraction. In *Proceedings of the Twelfth Meeting of the European Chapter of the Association for Computational Linguistics* (*EACL*), Athens, Greece, March–April 2009. Association for Computational Linguistics.

N. Verdezoto and L. Vieu. Towards semi-automatic methods for improving WordNet. In *Proceedings of the International Workshop on Computational Semantics* (*IWCS*), 2011.

L. von Ahn, M. Kedia, and M. Blum. Verbosity: A game for collecting common-sense facts. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 75–8, 2006.

K. von Fintel. Would you believe it? The king of France is back! Presuppositions and truth-value intuitions. In M. Reimer and A. Bezuidenhout, editors, *Descriptions and Beyond*. Oxford University Press, 2004.

Wikipedia. English Wikipedia snapshot, July 2009. URL en.wikipedia.org.

R. Wilensky. Knowledge acquisition and natural language processing. In A. Meyrowitz, editor, *Foundations of Knowledge Acquisition: Cognitive Models of Complex Learning*. Kluwer Academic Publishers, Boston, MA, 1993.

L. Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, 1953. Translated by G.E.M. Anscombe.

A Crowdsourced Evaluation & Filtering

A.1 Introduction

While it is expected that knowledge extraction systems will eventually produce sufficiently clean knowledge bases in order for inferences to be made about everyday things and events, currently the average quality of automatically acquired knowledge is not good enough to be used in traditional reasoning systems. An obstacle for knowledge extraction is the lack of an easy method for evaluating – and thus improving – the quality of results.

Evaluation in acquisition systems is typically done by human judging of random samples of output, usually by the reporting authors themselves (Lin & Pantel, 2002; Schubert & Tong, 2003; Banko *et al.*, 2007). This is time-consuming, and it has the potential for bias: it would be preferable to have people other than AI researchers label whether an output is commonsense knowledge or not. This chapter explores the use of Amazon's Mechanical Turk service, an online labor market, as a means of acquiring many non-expert judgements for little cost.

Previously, Snow *et al.* (2008) compared the quality of labels produced by non-expert Turkers against those made by experts for a variety of NLP tasks and found that they required only four responses per item to emulate expert annotations. Kittur *et al.* (2008) describe the use and necessity of verifiable questions in acquiring accurate ratings of Wikipedia articles from Mechanical Turk users. These results contribute to our methods below. While previous evaluations of Knext output have tried to judge the re-

lative quality of knowledge learned from different sources and by different techniques (Chapter 3), here the goal is simply to see whether the means of evaluation can be made to work reasonably, including at what scale it can be done for limited cost.

A.2 Experiments

A uniform random sample of factoids induced from the British National Corpus (BNC Consortium, 2001) was split into sets of 20. While obviously malformed results were removed, the more stringent filtering presented in Chapter 3 was omitted in order to ensure significant variation in the quality of the factoids to be rated.

The first evaluation followed the format of previous, offline ratings. For each factoid, Turkers were given the instructions and choices in Figure 3.3, rating on a scale of 1–5 with 1 being best. To help Turkers make such judgements, they were given a brief background statement:

> We're gathering the sort of everyday, commonsense knowledge an intelligent computer system should know. You're asked to rate several possible statements based on how well you think they meet this goal.

Mason & Watts (2009) suggest that while money may increase the number and speed of responses, other motivations, such as wanting to help with something worthwhile or interesting, are more likely to lead to high-quality responses. Participants were then shown the examples and explanations in Figure A.1. Note that while they are told some categories that bad factoids can fall into, the Turkers are not asked to make such classifications themselves, as this is a task where even experts have low agreement (Van Durme & Schubert, 2008).

Round 1 required participants to have a high (90%) approval rate. Under these conditions, out of 100 HITs[1], 60 were completed by participants whose IP addresses indicated

1 Human Intelligence Tasks – Mechanical Turk assignments. In this case, each HIT was a set of twenty factoids to be rated.

Examples of *good* statements:

1. *A song can be popular.*
2. *A person may have a head.*
3. *Maneuvers may be hold -ed in secret.*
   It's fine if verb conjugations are not attached or are a bit unnatural, *e.g.*, 'hold -ed' instead of 'held'.

Examples of *bad* statements:

1. *A thing may seek a way.*
   This is *too vague*. What sort of thing? A way for/to what?
2. *A cocktail party can be at Scotch Plains Country Club.*
   This is *too specific*. We want to know that a cocktail party can be at a country club, not at this particular one. The underscores are not a problem.
3. *A pig may fly.*
   This is *not literally true* even though it happens to be an expression.
4. *A word may mean.*
   This is *missing information*. What might a word mean?

Figure A.1: **The provided examples of good and bad factoids.**

they were in India, 38 from the United States, and 2 from Australia. The average Pearson correlation between the ratings of different Indian Turkers answering the same questions was a very weak 0.065, and between the Indian responders and those from the US and Australia was 0.132. On the other hand, the average correlation among non-Indian Turkers was 0.508 – close to the 0.6–0.8 range of correlations previously observed between experts, which can be taken as an upper bound on agreement for the task.

Given the sometimes subtle judgements of meaning required, being a native English speaker has previously been assumed to be a prerequisite. This difference in raters' agreements may thus be due to levels of language understanding, or perhaps to different levels of attentiveness to the task. However, it does not seem to be the case that the Indian respondents rushed: They took a median time of 201.5 seconds (249.18 avg. with a high standard deviation of 256.3 s – some took more than a minute per factoid). The non-Indian responders took a median time of just 115.5 s (124.5 avg., 49.2 std dev.).

Regardless of the cause, given these results, we restricted the availability of subsequent experiments to Turkers in the US. Ideally we would include other English-speaking countries, but there is no straightforward way to set multiple permitted countries on Mechanical Turk. Alternatively, a test of English comprehension could be posted with a satisfactory score as a prerequisite for participating in rating tasks, but this would significantly limit the number who would attempt the task. Even adding a location requirement significantly reduced the number of responses, with a sharp fall-off leading us to re-list the task with a higher pay-rate of 7¢ for 20 factoids *vs* 5¢ originally (Round 2).[2]

To avoid inaccurate ratings, we rejected submissions that were unreasonably quick or were strongly uncorrelated with other Turkers' responses. For each set of factoids (HIT), we collected five Turkers' ratings, and for each persons' set of responses computed the average of their three highest correlations with others' responses. We then rejected if the correlations were so low as to indicate random responses. The scores serve a second purpose of identifying a more trustworthy subset of the responses. (A cut-off score of

2 Rounds have been re-numbered from the original presentation of these results in Gordon *et al.* (2010a).

	All		*High Corr. (> 0.3)*		
Round	*Avg.*	*Std. Dev.*	*Avg.*	*Std. Dev.*	
1	2.59	1.55	2.71	1.64	BNC
2	2.80	1.66	2.83	1.68	BNC, US-only
3	2.61	1.64	2.62	1.64	BBC, simplified question
4	2.83	1.67	2.85	1.67	Weblogs, same format
5	2.75	1.64	2.75	1.64	Wikipedia, same format
6	2.76	1.61	2.89	1.68	BNC, coherent factoids only

Table A.1: **Average ratings from Mechanical Turk.** Lower numbers are more positive

0.3 was chosen based on hand-examination.) Table A.1 shows that these more strongly correlated responses rate factoids as slightly worse overall, possibly because those who are either casual or uncertain are more likely to judge favorably on the assumption that this is what the task authors would prefer. Or they may simply be more likely to select the top-most option, which was 'I agree'.

An example of a factoid that was labeled incorrectly by one of the filtered-out users is *A person may look at some thing-referred-to of press releases*, for which a Turker from Madras in Round 1 selected 'I agree'. Factoids containing the vague *thing-referred-to* are normally filtered out automatically, but leaving them in provided some obviously bad inputs for checking Turkers' responses. Another (US) Turker chose 'I agree' when told *Tes may have 1991es* but 'I disagree' when shown *A trip can be to a supermarket*.

We are interested not only in whether there is a general consensus to be found among the Turkers but also how that consensus correlates with the judgements of AI researchers. To this end, one of the authors rated five sets (100 factoids) presented in Round 2. The average correlation between all the Turkers and the author was 0.507, rising slightly to 0.532 if we only count those Turkers considered 'highly correlated' as described above.

As another test of agreement, for ten of the sets in Round 3, two factoids were designated as fix-points – the single best and worst factoid in the set, assigned ratings 1 and

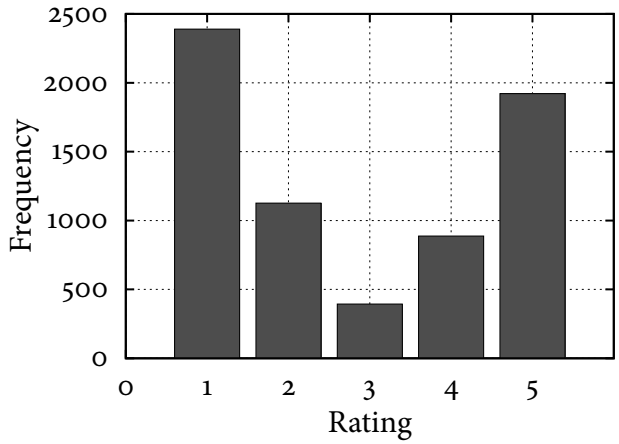

Figure A.2: **Frequency of ratings in the highly correlated results of Round 2.**

5 respectively. From the Turkers who rated these factoids, 65 of the 100 ratings matched the researchers' designations and 77 were within one point of the chosen rating. (If we only look at the highly correlated responses, this increases slightly to 68% exact match, 82% within one point.)

A few of the Turkers who participated had strong negative correlations to the other Turkers, suggesting that they misunderstood the task and were rating backwards.[3] Furthermore, one Turker commented that she was unsure whether the statement she was being asked to agree with (Figure 3.3) 'was a positive or negative'. To see how it would affect the results, we ran (as Round 3) twenty sets of factoids, asking simplified question 'Do you agree this is a good statement of general knowledge?' The choices were also reversed in order, running from 'I disagree' to 'I agree' and color-coded, with *agree* being green and *disagree* red. This corresponded to the coloring of the good and bad examples at the top of the page, which the Turkers were told to reread when they were halfway through the HIT. The average correlation for responses in Round 3 was 0.47, which is an improvement over the 0.34 avg. correlation of Round 2.

3 This was true for one Turker who completed many HITs, a problem that might be prevented by accepting/rejecting HITs as soon as all scores for that set of factoids were available rather than waiting for the entire experiment to finish.

Using the same format as Round 3, we ran factoids from two other corpora. Round 4 consisted of 300 random factoids taken from running Knext on weblog data and Round 5 300 random factoids taken from running Knext on Wikipedia. The average ratings for factoids from these sources are lower than for the BNC, reflecting the noisy nature of much writing on weblogs and the many overly specific or esoteric factoids learned from Wikipedia.

The results achieved can be quite sensitive to the display of the task. For instance, the frequency of ratings in Figure A.2 shows that Turkers tended toward the extremes: 'I agree' and 'I disagree' but rarely 'I'm not sure'. This option might have a negative connotation ('Waffling is undesirable') that another phrasing would not. As an alternative presentation of the task (Round 6), for 300 factoids, we asked Turkers to first decide whether a factoid was 'incoherent (not understandable)' and, otherwise, whether it was 'bad', 'not very good', 'so-so', 'not so bad', or 'good' commonsense knowledge. Turkers indicated factoids were incoherent 14% of the time, with a corresponding reduction in the number rated as 'bad', but no real increase in middle ratings. The average ratings for the 'coherent' factoids are in Table 3.2.

A.3 Conclusions

These initial experiments have shown that untrained Turkers evaluating the natural-language verbalizations of an open knowledge extraction system will generally give ratings that correlate strongly with those of AI researchers. Some simple methods were described to find those responses that are likely to be accurate. This work shows promise for cheap and quick means of measuring the quality of automatically constructed knowledge bases and thus improving the tools that create them. Beyond exploring the potential of Mechanical Turk as a mechanism for evaluating the output of Knext and other open knowledge extraction systems, these experiments also suggest the possibility of using crowdsourcing as a manual – but large-scale – filtering stage in the knowledge extraction process.

B Learning Lexical Axioms

WordNet provides a semantic hierarchy with broad lexical coverage, which has proved sufficiently precise to boost performance at many tasks involving natural language. However, it has not yet been formalized for use in a general reasoning system. In this chapter, we present such a formalization, designed to support inference with the knowledge learned from text in this dissertation. We use a semi-automatic annotation of WordNet with lexical features – most notably the mass–count distinction – to recognize inferentially different relations between concepts. The result is a collection of 77,263 lexical-semantic axioms, which are being released for general use. We evaluate a sample of the axioms for core concepts, showing their quality to be significantly better than a baseline interpretation of WordNet.

B.1 Introduction

We are interested in the creation of large knowledge bases to support language understanding and commonsense problem-solving. An important component of such a knowledge base is a large collection of lexical-semantic axioms relating more specific nominal concepts to more general ones. For example, axioms might assert that every rifle is a firearm, or that all malpractice is wrongful conduct. The most comprehensive, machine-readable source of this type of knowledge is WordNet (Fellbaum, 1998), which attempts to exhaustively enumerate and define the senses of each word. WordNet groups word senses considered synonymous into *synsets*, such as {firearm#1, piece#7, small-arm#1} or {wrongdoing#2, wrongful_conduct#1, misconduct#2, actus_reus#1}. These synsets are

linked by hierarchy relations: upward to a more general *hypernym* and downward to a more specific *hyponym*.[1]

These relations have been used to improve performance in natural language processing tasks such as information retrieval, document clustering, rule discovery, and text-based question-answering. These incremental improvements have not required that the knowledge used be sufficiently precise to support genuine language understanding, let alone commonsense reasoning. For these 'deep' problems, the knowledge needs to be regimented into some more precise, more reliable form. At the same time, it is desirable to keep a close connection between concepts referred to in ordinary language and their formalized versions. This way, the mapping from natural language to formalized representations (and vice versa) will be as direct and straightforward as possible in such applications as text-understanding and human–computer dialogue.

As such, it is natural to ask whether WordNet senses of nouns can be used directly as predicates in a formalized knowledge base, with hierarchy relations (in the upward direction) corresponding to universally quantified conditional (*i.e.*, if–then) formulas. For example, can we make the formal claim that, for appropriate senses, every rifle is a firearm, or that all malpractice is wrongful conduct, as in the following?

$$\forall x.\, \text{rifle1.n}(x) \Rightarrow \text{firearm1.n}(x),$$
$$\forall x.\, \text{malpractice1.n}(x) \Rightarrow \text{wrongful\_conduct1.n}(x)$$

For people to judge these claims, we want to verbalize them as corresponding English statements:

Every rifle is a firearm.

Every amount of malpractice is an amount of wrongful conduct.

1 WordNet also includes antonymy, part–whole, and membership relations among others. In this section I focus on nominal hypernyms, WordNet's most extensive and most used component.

The difference in these phrasings reflects a difference in the meaning of the terms involved: Rifles are distinct, individuable entities; malpractice is less discrete. More generally, while mass terms are cumulative, count terms are not (Pelletier & Schubert, 1989). A test phrasing highlights the difference:

> Some malpractice and some more malpractice constitute an amount of malpractice.
>
> * Some rifle and some more rifle constitute an amount of rifle.

Consistent with the close semantic connection between mass terms and plurals (Nicolas, 2014), if we change 'rifle' to 'rifles', the claim becomes true, even if speakers disprefer the phrasing 'amount of' applied to a plural.

For other hyponym–hypernym pairs, different relations are appropriate. For instance, the count noun 'plank' and the mass noun 'lumber' are related as *Every plank is an amount of lumber*. When a hyponym denotes an individual or a generic kind, the appropriate relation is not subsumption but instantiation, *e.g.*, *Gold is a noble metal*. We sort out this ambiguity by looking at logical test phrasings and considering the criteria for detecting these meanings, most notably the mass–count distinction. In §B.3, we enumerate the relations we find between synset members in WordNet's hypernym hierarchy and the conditions under which they hold, but we first consider whether some alternative lexical and ontological resources would be less problematic and whether previous work in refining WordNet could help.

B.2 Previous Work

One of the most noteworthy efforts in knowledge engineering for artificial intelligence is the Cyc project (Lenat, 1995), which has created a collection of world knowledge in the CycL logical form, including a manually constructed core ontology of over 200,000 terms. However, Cyc lacks the systematic link to language we find in WordNet, as seen

in constructed predicates like *CoexistingWithSomethingElse* or *OrganismByTaxonomic-Kingdom-Biology-Topic*.

Another significant resource is SUMO (Niles & Pease, 2001), a formal ontology consisting of around 80,000 axioms, including information about classes and their meanings. Álvez *et al*. (2012) translated most of its 1,000-term upper ontology into first-order logic (FOL) and demonstrated its use for commonsense inference. However, while the logical formulation of axioms in SUMO is appealing for reasoning, it does not meet our goals of lexical and conceptual coverage. SUMO maps WordNet synsets to its own formal terms for broader coverage, but these mappings are coarse and lose the specific meaning of the synset. For instance, in WordNet, *stage dancing* (*e.g.*, ballet) has the hypernyms *dancing* (the act) and *performing arts* (the discipline). In SUMO, 'stage dancing' is mapped to the class *Dancing*, but there is no sense for this as a discipline or an art, only as *Body-Motion*.

Pustejovsky (1995) presents the alternative approach of a 'generative lexicon', recognizing that in different contexts a word will express different meanings, making it infeasible to try (as WordNet does) to enumerate them independently of context. Rather, he argues, a lexical entry should provide the information necessary to derive the sense the word will take on in a given context. Resulting work on the Brandeis Semantic Ontology (Pustejovsky *et al*., 2006) may eventually provide a more consistent basis for lexical axioms, but no resource has yet been released.

For WordNet, Kaplan & Schubert (2001) previously looked at the accuracy of taking the noun hierarchy as a simple subsumption taxonomy. They identified a number of the problems we address in this chapter, including the conflation of individuals, predicates, and kinds and the mixing of mass and count uses of terms. More recently, Verdezoto & Vieu (2011) presented promising work to automatically identify problematic relations in WordNet based on conflicts between meronyms and hyponyms. However, neither effort attempted to produce a corrected formal resource as we do here.

Several lines of previous work have sought to address 'is-a' ambiguity by making WordNet a formal ontology, manually separating or removing non-subsumptive hypernym relations and restructuring the upper, most abstract levels of the hierarchy to fit different ontological principles. Most notable is the work of Nicola Guarino and his collaborators (*e.g.*, Guarino & Welty, 2000) on distinctions and design principles for producing cleaner ontologies. This resulted in the construction of the DOLCE upper-level ontology and the alignment of WordNet (1.6) subtrees to it, to make OntoWordNet (Gangemi *et al.*, 2003).

In contrast with these lines of work, we are less concerned with ontological hygiene than with inferential efficacy for intuitively plausible reasoning and understanding. Rather than restrict the content of WordNet to its subsumptive relations, we automatically produce lexical axioms that formalize a variety of logical relations between synset members. To the best of our knowledge, this is the first such attempt.

B.3 Acquiring Axioms

No work has yet provided a large-scale set of reasonably reliable lexical axioms that are closely integrated with language. To address this problem, we studied random samples of WordNet hypernym relations, formulated appropriate logical axioms, and then formed hypotheses about what features indicate that a relationship holds. In this section, we present the types of relations we find in WordNet, the method of generating axioms, and the criteria we use.

B.3.1 *Distinguishing Relations*

In Table B.1, we give the frequency of axioms resulting from the following schemata, both in WordNet as a whole and in the 'core' axioms used for evaluation (see §B.4). The axiom schemata are presented in conceptual groups, which are used to balance the evaluation sample.

Group 1: Count Nominals

1 *Every ϕ is a ψ.*

($\forall x$: [x ϕ.n] [x ψ.n])

ϕ and ψ are singular count nominals. The hypernym relation holds for each individual, *e.g.*,

> *Every cat is a feline.*
> ($\forall x$: [x cat1.n] [x feline1.n])

Group 2: Mass Nominals and Plurals

2 *Every amount of ϕ is an amount of ψ.*

($\forall x$: [x ϕ.n] [x ψ.n])

ϕ and ψ are either mass terms ('water') or lexical plurals ('cattle'). *E.g.*,

> *Every amount of red wine is an amount of wine.*
> ($\forall x$: [x red_wine1.n] [x wine1.n])

This shares the logical form of Schema 1 but differs in its verbalization. As discussed in the introduction, this is an indication that singular count nouns are inferentially distinct from plurals and mass terms with respect to cumulativity. See the inference example in §B.5.

Group 3: Kinds and Individuals

3 *(The) ϕ is/are a ψ.*

[ϕ.name ψ.n]

ϕ is an individual – either an individual name ('Belgium') or a name-like designation ('Homo sapiens') for a *generic kind* (Carlson, 1977a,b). ψ is a singular count nominal. *E.g.*,

> *AIDS is an immunodeficiency.*
> [AIDS.name immunodeficiency1.n]

4 *(The) ϕ is/are a ψ.*

[(k ϕ.n) ψ.n]

ϕ is a mass nominal ('oil') or a count nominal known to function as a natural kind ('tiger'). ψ is a singular count nominal that is a kind-level predicate ('species'). This schema gives claims about generic kinds, formed with EL's kind reification operator 'k'. *E.g.*,

> *Gold is a noble metal.*
> [(k gold3.n) noble_metal1.n]

5 *Every ϕ is an item of ψ.*

($\forall x$: [x ϕ.n] [x item-of.n (k ψ.n)])

ϕ is a singular count nominal. ψ is an *atomic ensemble*, a mass term that cannot be arbitrarily subdivided ('furniture', not 'water'); see §B.3.3. Equivalently, ψ can be a plural; the atomic ensembles denoted by plurals are in no way logically distinguishable from atomic ensembles denoted by mass terms. *E.g.*,

> *Every bomb is an item of weaponry.*
> ($\forall x$: [x bomb1.n] [x item-of.n (k weaponry1.n)])

6 *(The) ϕ is/are a branch of (the) ψ.*

[(k ϕ.n) branch-of.n (k ψ.n)]

ϕ and ψ are fields of study. These are hyponym descendants of discipline#1, excluding kind-level predicates like humanistic_discipline#1. *E.g.*,

> *Astronomy is a branch of physics.*
> [(k astronomy1.n) branch-of.n (k physics1.n)]

Group 4: Transitional

7 *Every ϕ is an amount of ψ.*

($\forall x$: [x ϕ.n] [x ψ.n])

ϕ is a singular count nominal and ψ is a plural or a mass noun. *E.g.*,

Every document is an amount of written material.

($\forall x$: [x document1.n] [x written_material1.n])

8 *Every amount of ϕ is a ψ.*

($\forall x$: [x ϕ.n] [x ψ.n])

ϕ is a mass term or a plural that is a hyponym of group#1 or measure#2. ψ is a singular count nominal. *E.g.*,

Every amount of people is a group.

($\forall x$: [x people1.n] [x group1.n])

9 *Every amount of ϕ is an amount of ψs.*

($\forall x$: [x ϕ.n] [x (plur ψ.n)])

ϕ is a mass term or a plural count nominal, and ψ is a singular, object-level predicate. *E.g.*,

Every amount of baggage is an amount of cases.

($\forall x$: [x baggage1.n] [x (plur case5.n)])

Group 5: Events

10 *Every ϕ is a ψ.*

($\forall x$: [x (plur ϕ-c.n)] [x (plur ψ-c.n)])

ϕ and ψ are events that have both a mass and a count sense. Here the 'plur' operator is 'massifying' the count (discrete-only) sense of an event predicate to match possible iteration. *E.g.*,

Every restoration is a repair.

($\forall x$: [x (plur restoration2-c.n)] [x (plur repair1-c.n)])

11 *Every ϕ is a ψ.*

($\forall x$: [x (plur ϕ-c.n)] [x ψ.n])

ϕ and ψ are events. ϕ is ambiguous between mass and count senses, while ψ is count. *E.g.*,

Every sinning is a transgression.

($\forall x$: [x (plur sinning1-c.n)] [x transgression1.n])

12 *Every ϕ is a ψ.*

($\forall x$: [x ϕ.n] [x (plur ψ-c.n)])

ϕ is count and ψ is ambiguous between mass and count senses. *E.g.*,

Every dance step is a locomotion.

($\forall x$: [x dance_step1.n] [x locomotion2-c.n])

B.3.2 *Method*

While the members of each synset are closely related, we found they are not always interchangeable as predicates. Many synsets contain a mix of mass and count, singular and plural, *e.g.*, {cutlery#2, eating_utensil#1}. An eating utensil is an *item of* cutlery. Thus, in relating the synset to its hypernym, tableware#1, we form separate axioms for both of these predicates. However, making an axiom for every combination of word senses in a pair of synsets would lead to an unnecessary explosion in the number of axioms. Instead, those members that share logically equivalent properties (mass terms, atomic ensembles & lexical plurals, singular count nouns, individual names) can be stated to be synonymous and an axiom can use a single representative predicate.

Thus, our method is: For each hyponym–hypernym pair, select the synset members with distinct properties for which we will form axioms. Then for the Cartesian product of the selection sets, check each pair of word senses against the restrictions for each schema and output an axiom when they match – see Figure B.1.

B.3.3 *Determining the Mass–Count Distinction and Other Significant Features*

Mass–Count Annotating WordNet with mass and count information requires not just determining if a word is usually mass or count but, for many ambiguous words, whether

$\mathcal{S}$, the set of nominal synsets in WordNet
$\mathcal{H}$, the set of all hyponym–hypernym synset pairs
$\mathcal{C}$, the set of all annotation categories.

```
GENERATE-AXIOMS(H):
    for synset pair ⟨P, Q⟩ ∈ H:
        L ← SELECT-MEMBERS(P)
        L′ ← SELECT-MEMBERS(Q)
        for lemma pair ⟨φ, ψ⟩ ∈ L × L′:
            for each schema s:
                if ⟨φ, ψ⟩ matches the argument restrictions of s:
                    Instantiate s with ⟨φ, ψ⟩

SELECT-MEMBERS(s ∈ S):
    R ← {}
    for lemma l ∈ s:
        for category c ∈ C:
            if l is annotated as c:
                f ← False
                for r ∈ R:
                    if r is annotated as c:
                        if l has a lower sense number than r:
                            Replace r with l in R
                        f ← True
                        break
                    if not f:
                        R ← R ∪ {l}
    return R
```

Figure B.1: **Algorithm for axiomatizing WordNet's hypernym hierarchy.**

Group	*Schema*	*All Axioms*	*Core Axioms*
G1	**1**	47,450	4,397
G2	**2**	7,114	979
G3	**3**	2,838	46
	4	1,998	497
	5	887	119
	6	637	23
G4	**7**	5,233	578
	8	187	23
	9	4,757	430
G5	**10**	1,892	649
	11	1,606	452
	12	2,664	553
	All	77,263	8,746

Table B.1: **Axiom counts for WordNet as a whole and 'core' synsets.** The schemata are described in §B.3.1. The 'core axioms' are those from which the evaluation set was sampled, described in §B.4.

a particular word sense is. While sometimes WordNet splits a word into different synsets for its mass and count senses, in other cases it coerces a single synset's meaning through multiple hypernym links. For instance, coffee#1 is both a liquid#1 (mass) and a beverage#1 (count), reflecting different uses:

'How much coffee did you drink?'

'I'll have a coffee.'

Often we consider one form to be basic and the other derived. While coffee is primarily mass, we understand a count use to mean a standard portion of it, *i.e.*, a cup. (See comments in §B.7.)

Various past studies have been aimed at classifying lexemes as mass, count, or both, (*e.g.*, Bond & Vatikiotis-Bateson, 2002; Baldwin & Bond, 2003; O'Hara *et al.*, 2003; Lapata & Keller, 2004, 2005). Typically these have used multiple sources of information, such as morphology, corpus occurrence environments, the Cyc knowledge base, and seemingly similar lexemes in WordNet. While Álvez *et al.* (2008) semi-automatically annotated WordNet 1.6 with EuroWordNet's Top Concept ontology semantic features, including *Substance* and *Object* – rough analogues of mass and count – we found these annotations too noisy, *e.g.*, labeling *cytostome* (a cell mouth) a substance. Most recently, Kiss *et al.* (2014) conducted the manual annotation of WordNet senses with more fine-grained distinctions based on test sentence phrasings. *E.g.*, they annotate fruitcake#1 as a *dual life* noun where the same WordNet sense allows both count readings and mass readings but whiskey#1 is taken as *uncountable with sorter/packager plurals* meaning that it is inherently mass but allows the specific count use for 'a [glass of] whiskey', while seawater#1 is fully uncountable. These results were not available when our axiomatization was done but may be used to improve future classification.

For this work, we first annotate each general noun sense in WordNet 3.1 as a plural or singular count term, an atomic ensemble, or a non-atomic mass term. To do so, we use two sources of information: syntactic patterns and existing dictionaries. For each noun, we search the Google n-grams data set (Brants & Franz, 2006) for occurrences in

mass and count syntactic patterns based on those given by Bunt (1985). *E.g.*, 'many', 'few', 'fewer', 'fewest', 'several', or 'numerous' *x*s are indicative of a count use of *x* while 'much' or 'little bit of' *x* are indicative of a mass use. The n-gram data allows broader lexical coverage than matching against traditional text corpora, but it provides less context and the patterns yield only moderate accuracy. *E.g.*, while 'a *x*' typically indicates that *x* has a count sense, the pattern can erroneously match references such as 'grade A milk'. We supplement this classification by looking up each lemma online in the Oxford Dictionary (http://oxforddictionaries.com) and Wiktionary (http://wiktionary.org) and counting the number of senses marked as being count or mass (uncountable).

When, based on this information, a lemma is ambiguous between mass and count readings, the lemma is labeled by a decision tree that checks features of the lemma and the synset, including:

1 lemma name and morphology: *E.g.*, beginning with 'period of' or 'piece of', (count); ending in 'powder', 'oil', '-ness', or '-ity' (mass).
2 gloss: *E.g.*, beginning with 'material' (mass).
3 hypernym ancestors: *E.g.*, *artifact*#1 (count); *chemical element*#1 (mass).
4 hyponym glosses: *E.g.*, a hyponym of *x* is defined as 'a/an *x* that…' (count); a hyponym of *x* is defined as 'an amount/quantity/portion of *x* (mass).
5 example sentences: *E.g.*, an example including 'a/an *x*' (count).

If the word is ambiguous but the other members of the synset are all known to be mass or to be count, then the lemma is labeled the same.

Atomic Ensembles In axiom schema 5, we also need to recognize those mass predicates that apply to *atomic ensembles* (also called *aggregate terms*). Unless we're thinking scientifically, some mass nouns can be divided arbitrarily into more of the same: All *air* has proper parts that are also air, all *meat* has proper parts that are also meat, etc. On the other hand, it is not the case that all *poultry*, *furniture*, *foliage*, *cutlery*, or *dinnerware* has proper parts that are also poultry, furniture, etc. respectively. Rather, these mass terms

denote entities that have atomic parts. *E.g.*, a chair is furniture, but it has no proper parts that are also furniture. We are not aware of any attempt to enumerate atomic ensembles in English or to automatically find them in text. For this work, we annotated 178 WordNet lemmas as atomic ensembles with reference to the linguistic literature. An additional 915 lemmas are identified (based on dictionary entries) as lexical plurals. As we noted for axiom schema 5, plurals also denote atomic ensembles, which are logically indistinguishable from mass ones.

Individuals It is also important for us to distinguish individuals (*e.g.*, 'Nikola Tesla') from common nouns. Following the criticism of ontologists like Aldo Gangemi, WordNet 2.1 began to move instances from hyponym relations to instantiation relations (Miller & Hristea, 2006). In WordNet 3.1, there are 15,675 such word senses. These give us a source for formulas about many important events, people, states, and other kinds of individuals. Identifying individuals also lets us avoid nonsensical quantification over 'Every Nikola Tesla'. However, we find that many individuals are still mingled with classes as hyponyms, *e.g.*, 'The Industrial Workers of the World' (a specific union) or 'St Polycarp' (a specific martyr). We identified 13,561 additional individuals by checking whether Wiktionary only lists a lemma only as a proper noun and by manually inspecting synsets where all lemmas are capitalized.

B.4 Evaluation

Like many efforts in knowledge acquisition and reasoning, the creation of lexical semantic axioms is motivated by a variety of applications, but it is not easily evaluated through them. Instead, it is traditional to rely on human judgements – often those of the authors – to determine the accuracy and appropriateness of the results (as in Friedland & Allen, 2003; Banko *et al.*, 2007; Van Durme & Schubert, 2008; Carlson *et al.*, 2010).

While it is natural for us to judge a random sample of the resulting axioms, this would not accurately reflect their value for tasks requiring commonsense reasoning. Due

to WordNet's broad lexical coverage, most of the words it includes are rare, including, for instance, specialized scientific and medical terminology. Therefore, as our evaluation set, we took the axioms where both predicates are from a set of 'core' synsets. These are the union of two standard lists: Boyd-Graber *et al.*'s Core WordNet (2006) and Izquierdo *et al.*'s Base Concepts (2007). To get a balanced sample of different relations, we randomly selected 40 axioms for each of the five schemata groups presented in § B.3. For rating, these axioms were shuffled with a baseline interpretation of the same hyponym–hypernym pairs, corresponding to Schema 1, the most common. A random selection of the English verbalizations of our output and baseline output from the evaluation set is presented in Figure B.2 and Figure B.3.

Judges were asked to evaluate each axiom's English verbalization based on whether it is a reasonable claim with respect to the general word senses indicated by the definition and examples for the synset. They were instructed that whenever an axiom says 'amount of' they should apply the cumulativity test described in the introduction: Does 'some x and some more x' constitute 'an amount of x'? They were asked to apply the same test to 'a(n)' and 'every' phrasings; while 'amount of' applied to plurals should be tolerated, 'every' or 'a(n)' applied to mass terms should not.

Each axiom was rated on a scale of 1 (best) to 5 (worst). The authors each rated the full evaluation set of 400 axioms. The 200 axioms of system output had an average rating of 1.78, while the 200 baseline axioms had an average rating of 3.29. A Pearson correlation of .77 reflects a high level of agreement.

For greater objectivity, it is desirable to also have judges unaffiliated with the work rate the axioms. However, we found it difficult to train judges to be sufficiently sensitive to the property of cumulativity and to resist type-ifying the claims to allow non-basic readings such as 'a wine is a liquid'. One judge's ratings for 200 axioms (100 system output, 100 baseline) were well-correlated with the authors (0.6), giving our output an average rating of 1.71 and the baseline 2.59. A second judge's ratings were less well-correlated

	Schemata Groups						
Rating	G1	G2	G3	G4	G5	All	*Baseline*
Best **1**	93	85	103	67	51	399	211
2	11	18	8	19	21	77	54
3	6	9	6	12	15	48	46
4	7	5	1	12	20	45	131
Worst **5**	3	3	2	10	13	31	158
Average	1.47	1.53	1.21	1.99	2.36	1.72	2.95

Table B.2: **Distribution of ratings for WordNet axioms.** The evaluation set is 200 axioms of system output and 200 of the baseline, each rated by three judges. Ratings of system output are broken down by the schemata groups from §B.3.1.

(0.45), indicating difficulty in understanding the criteria or in assessing them, but still rated our system's output better on average (1.47) than the baseline (1.97).

The distribution of ratings for all three judges (200 sys. ratings by each of the authors, 100 by each of the other judges; likewise for the baseline) can be seen in Table B.2, including a breakdown of the ratings by the axiom schemata groups, showing their relative reliability.

B.5 Reasoning with WordNet Axioms

The need for knowledge about entailment relations between entity types has been recognized since the early days of AI (*e.g.*, Amsler & White, 1979; Chodorow *et al.*, 1985; Wilensky, 1993). To make commonsense inferences, it is especially important to have the sort of taxonomic knowledge contained in WordNet's hypernym hierarchy. For example, if we are told 'Merry is a cat', a basic reasoning chain is: *Every cat is a feline, every feline is a carnivore, . . . , every chordate is an animal.* Therefore, *Merry is an animal.* This process of generalization allows us to apply world knowledge known at a higher level of generality. For instance, if we know *Every animal needs food to live*, we can conclude *Merry*

Every amount of reparation is an amount of compensation.
(∀ *x*: [*x* reparation1.n] [*x* compensation1.n])
— *reparation#1*: compensation (given or received) for an insult or injury
— *compensation#1*: something (such as money) given or received as payment or reparation (as for a service or loss or injury)

Curiosity is a cognitive state.
[(k curiosity1.n) cognitive_state1.n]
— *curiosity#1*: a state in which you want to learn more about something
— *cognitive_state#1*: the state of a person's cognitive processes

Every cathedral is a church building.
(∀ *x*: [*x* cathedral1.n] [*x* church_building1.n])
— *cathedral#1*: any large and important church
— *church_building#1*: a place for public (especially Christian) worship

Every abandonment is a rejection.
(∀ *x*: [*x* (plur abandonment1-c.n)] [*x* (plur rejection1-c.n)])
— *abandonment#1*: the act of giving something up
— *rejection#1*: the act of rejecting something

Every assembly is a gathering.
(∀ *x*: [*x* assembly4.n] [*x* gathering1.n])
— *assembly#4*: a group of persons who are gathered together for a common purpose
— *gathering#1*: a group of persons together in one place

Every counting is an investigation.
(∀ *x*: [*x* counting1.n] [*x* (plur investigating1-c.n)])
— *counting#1*: the act of counting; reciting numbers in ascending order
— *investigation#2*: the work of inquiring into something thoroughly and systematically

Computer science is a branch of applied science.
[(k computer_science1.n) branch-of.n (k applied_science1.n)]
— *computer_science#1*: the branch of engineering science that studies (with the aid of computers) computable processes and structures
— *applied_science#1*: the discipline dealing with the art or science of applying scientific knowledge to practical problems

Figure B.2: **System output from the evaluation set.**

Every restoration is a repair.
(∀ *x*: [*x* restoration2.n] [*x* repair1.n])
— *restoration#2*: the act of restoring something or someone to a satisfactory state
— *repair#1*: the act of putting something in working order again

Every physics is a natural science.
(∀ *x*: [*x* physics1.n] [*x* natural_science1.n])
— *physics#1*: the science of matter and energy and their interactions
— *natural_science#1*: the sciences involved in the study of the physical world and its phenomena

Every pretending is a dissimulation.
(∀ *x*: [*x* pretending1.n] [*x* dissimulation1.n])
— *pretending#1*: the act of giving a false appearance
— *dissimulation#1*: the act of deceiving

Every sameness is a quality.
(∀ *x*: [*x* sameness1.n] [*x* quality1.n])
— *sameness#1*: the quality of being alike
— *quality#1*: an essential and distinguishing attribute of something or someone

Every solid is a matter.
(∀ *x*: [*x* solid1.n] [*x* matter3.n])
— *solid#1*: matter that is solid at room temperature and pressure
— *matter#3*: that which has mass and occupies space

Every encroachment is an influence.
(∀ *x*: [*x* encroachment3.n] [*x* influence2.n])
— *encroachment#3*: influencing strongly
— *influence#2*: causing something without any direct or apparent effort

Every sand is a dirt.
(∀ *x*: [*x* sand1.n] [*x* dirt1.n])
— *sand#1*: a loose material consisting of grains of rock or coral
— *dirt#1*: the part of the earth's surface consisting of humus and disintegrated rock

Figure B.3: **Baseline output from the evaluation set.**

needs food to live, and an intelligent agent might, accordingly, form the goal of feeding her.

A slightly more complex line of reasoning demonstrates the inferential importance of the semantic distinctions we have explored in this chapter:

All gold dust is gold.
(∀ *x*: [*x* gold_dust1.n] [*x* gold3.n])

Gold is a noble metal.
[(k gold3.n) noble_metal1.n]

A meta-axiom over mass predicates gives the logical equivalence of our 'amount of' verbalizations:

All p *is an amount of the kind* p *(for mass predicate* p*).*
($\forall_{\text{pred}}$ *p*: ['*p* mass-pred]
(all *x* [[*x p*] ⇔ [*x* amount-of (k *p*)]]))

And, from our annotation of WordNet, we know

Gold_dust1.n *and* gold1.n *are mass predicates.*
['gold_dust1.n mass-pred], ['gold3.n mass-pred]

Therefore,

Every amount of gold dust is an amount of a (certain)
noble metal.
(∃ *y*: [*y* noble_metal1.n]
(∀ *x*: [*x* amount-of (k gold_dust1.n)]
[*x* amount-of *y*]))

(Rather than *All gold dust is noble metal* or *Every gold dust is a noble metal*!) So if, ignoring tense, we learn

John found some gold dust.
(∃*x*: [*x* gold_dust1.n] [John.name find.v *x*])

We can conclude

John found some amount of a noble metal.
(∃*y*: [*y* noble_metal1.n]
(∃*x*: [*x* amount-of *y*]
[John.name find.v *x*]))

B.6 Conclusions

We have seen that WordNet's hypernym hierarchy represents a variety of semantically distinct relations. To create lexical axioms suitable for use in a general reasoner, we must identify and formalize these relations. In this chapter, we've shown that we can use the mass–count distinction to obtain a large number of such axioms, which are judged significantly better than a subsumptive count-noun baseline.

B.7 Discussion and Future Work

Our results show that we can significantly improve the reliability of hierarchy axioms extracted from WordNet by attending to the mass–count distinctions among word senses (and some other subtle properties). But our research undertaking has also revealed some systematic difficulties in making logical sense of WordNet hierarchy relations, and these point to interesting possibilities for future work.

Consider this hyponym–hypernym pair:

watching#1: the act of observing; taking a patient look

looking_at#1: the act of directing the eyes toward something and perceiving it visually

Both glosses characterize the word senses in terms of acts, and since 'act' is a count noun (both syntactically and conceptually – there can be single acts or multiple, distinct acts), one would expect favorable judgements for an axiom expressing *Every watching is a looking-at*. However, the oddity of the paraphrase *All watchings are lookings-at* makes the count readings of these nouns rather suspect. In fact, natural occurrences like 'his watching you while you sleep' (a gerund) or 'his watching of CNN' (a deverbal noun) suggest that watching is basically an activity rather than an act. However, this is an elusive intuition as we can easily conceive of bounded *episodes* of any activity, which have the character of acts. Indeed, we can assume that there is a class of 'countifying' operators that map activity/process predicates to action/event predicates; for example, adverbials such as 'for three hours' accomplish such a transformation (see Hwang & Schubert, 1994), which systematically treats the semantics of durative and many other types of adverbials; the analysis is for verb phrase adjuncts, but many of the observations carry over to deverbal nouns.)

The relationship between the above two word senses is further obscured by the fact that the synset for 'watching#1' also contains the word sense 'observation#2', for which the definition as an act seems to fit better: We can naturally speak in the plural of 'Penn's observations of the nightly newscasts from Vietnam', and since the newscasts are bounded events, so are the observations.

This is just one example of the type-shifting (countability-shifting, and, for deverbal nouns, aspectual-category-shifting) transformations that many nouns are susceptible to, and how this shifting potential can confound intuitive judgements of the validity of hierarchy axioms derived from WordNet synsets. Here are some more shifting operations:

1 *Iteration*: *e.g.*, the event predicate 'sneeze' becomes a predicate true of the activity of 'sneezing' through iteration; WordNet places 'sneeze#1' and 'sneezing#1' in the same synset.
2 *Grinding*: *e.g.*, while in nature we find 'potatoes', when eating we may have some amount of [mashed] potato. This reading of '*x* stuff' is what Pelletier (1975) called the

universal grinder. WordNet distinguishes potato#1, the edible root, from potato#2, the entire plant (*i.e.*, a potato vine). However, that leaves both count and mass readings of the first sense. It is a long-standing question whether these readings result from coercion of a single sense or are distinct senses.

3 *Raising to a kind-level predicate*: *e.g.*, the predicate, 'wine', true of quantities of stuff becomes a kind of wine (true, for example, of Merlot or Riesling); WordNet does not have an entry for the latter sense; on the other hand, WordNet's entry 'medicine#2' is grouped into the same synset with 'drug#1' – clearly a *kind* of medicine. (This is what Bunt (1985) called the *universal sorter* – the 'kind of *x*' reading.)

4 *Conventional portions*: 'a wine' or 'a beer' can refer to a serving of either stuff, perhaps derivable from the basic predicate by a 'conventional-portion-of' or 'serving-of' operator. (This is what Jackendoff (1991) called the *universal packager*.) WordNet does not distinguish these senses, but its second entry for 'tissue', the synset {tissue#2, tissue_paper#1}, is accompanied by a gloss that is compatible with either a predicate true of any amount of tissue paper, of certain standard portions ('Please hand me a tissue'), or of certain kinds of paper. Similarly, the gloss for 'physical exercise' describes this as activity (thus mass), but the synset also contains 'workout', which clearly refers to a conventional bout of exercise.

These observations suggest that future work should look into systematic generation of meaning variants from certain 'basic' meanings of nouns. WordNet synsets would then be analyzed (as far as possible automatically) to identify and relate meaning variants within synsets, generated by type-shifting operators of the above types. (The total number of such operators appears to be quite small.) Such a project would fit well with Pustejovsky's Generative Lexicon project (Pustejovsky, 1995), and would enable a more complete and accurate axiomatization of WordNet hierarchy relations. Of course, complete reliability is unattainable, if only because WordNet is not error-free. For example, WordNet relates 'identity#2' to hypernym 'recognition#2', but the former is defined in terms of individual characteristics and the latter as a process, and there is no way in

which characteristics can be construed as a process. But these examples seem to be relatively rare (based on informal observation, perhaps a few out of 100), so there remains considerable scope for extracting relatively reliable, more refined formalized knowledge from WordNet.