Notes on the Acquisition of Conditional Knowledge

Benjamin Van Durme[1]
Department of Computer Science
University of Rochester
Rochester, NY 14627

June 2008
Technical Report 937

[1]Supported by NSF grants IIS-0328849 and IIS-0535105. 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.

Abstract

Research in Information Extraction has been overly focused on the extraction of facts concerning individuals as compared to general knowledge pertaining to classes of entities and events. In addition, preference has been given to simple techniques in order to enable high volume throughput.

In what follows we give examples of existing work in the field of knowledge acquisition, then follow with ideas on areas for exploration beyond the current state of the art, specifically with respect to the extraction of conditional knowledge, making use of deeper linguistic analysis than is currently the norm.

Contents

Chapter 1

Introduction

Within the field of *Artificial Intelligence (AI)*, the so-called *knowledge bottleneck* has long hampered progress towards the goal of thinking, speaking machines. Over the last twenty years, due to the rapidly increasing amount of available data, not to mention computing power and storage, a new discipline has emerged aimed at solving this bottleneck. *Information Extraction (IE)* is an area of research spanning *Information Retrieval*, *Machine Learning*, *Databases*, *Knowledge Representation* and *Natural Language Processing.* IE researchers seek to develop algorithms and textual resources for acquiring large amounts of information automatically. To date, the majority of the community has focused on *fact extraction*, which looks for information about specific individuals, locations, dates, times, etc. This sort of knowledge is useful for tasks such as answering simple questions, e.g., *Who invented the traffic cone?* (Voorhees and Tice, 2000).

Our interests lie elsewhere, in the acquisition of general sorts of knowledge about how the world operates. For example, rather than being able to fill the empty slot in *?x invented the traffic cone*, we want to know that *People may invent things.*

Examples of specific topics in *Knowledge Acquisition (KA)* include the discovery of: hypernym relations, e.g., *Bambara ndang isa bow lute* (Hearst, 1992); general propositions, e.g., *Children may live with relatives* (Schubert and Tong, 2003); characteristic attributes of concept classes, e.g., *Countries may have presidents* (Paşca and Van Durme, 2007); paraphrase rules, e.g., *X wrote Y if and only if X is the author of Y* (Lin and Pantel, 2001a); and common verb-verb relations, e.g., *buy happens-before sell* (Chklovski and Pantel, 2004).

More specifically, we aim to supplement current techniques with methods that are less dependent on redundancy in favor of a richer initial extraction process. By placing increased priority on natural language semantics within KA, we intend to enable the discovery of knowledge that is both more complex and less commonly expressed, even within large corpora.

In the following two sections we give an overview of some of the existing knowledge resources as well as prior work related to commonsense knowledge acquisition. This is followed by a description of our own work in data driven extraction. Finally we turn to thoughts on the extraction of conditional knowledge.

Chapter 2

Existing Knowledge Resources

Our long term goal is a comprehensive knowledge base enabling useful, automated inference over a variety of everyday topics.

There are a variety of existing repositories that may at least partially satisfy this need. Here we give mention to a few of the most well known: WordNet, along with its later derivatives; and the Cyc knowledge base.

2.1 WordNet

From the WordNet website:[1]

> WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets.

WordNet is perhaps the most well known and widely used lexical resource within the computational linguistics community. For instance, its network structure is widely used as a basis for ad hoc similarity scores based on path distances[2], enabling a variety of projects substituting similarity for richer world knowledge, such as our own work in *Question Answering (QA)* (Van Durme et al., 2003).

[1] `http://wordnet.princeton.edu/`

[2] A collection of references to such systems is maintained is by Pedersen (2007).

Most relevant for us here is the use of WordNet as a formal ontology. By interpreting hypernym links as classically understood *is-a* relations, nodes that had originally been intended to represent sets of synonymous lexical terms may alternatively be considered as standing for concept classes.

That there exists a distinction between these interpretations and that such a distinction could be meaningful, has been been discussed by a number of authors. Exploring the use of WordNet within our group's reasoning system, Kaplan and Schubert (2001) concluded that the notions of *is-a* and *is-a-kind-of* are impossibly confused under the single notion of hypernym. [3]

A recent release of WordNet saw a subset of hypernym links renamed as *instance-of* (Miller and Hristea, 2006). Unfortunately their notion of what comprises an instance relation is more limited than what we would hope. For example, languages, such as *Italian* or *Spanish*, are not instances, while sacred texts such as *Bhagavadgita* or *Zend Vesta* are. *Bible* is not, but its various versions such as *King James* or *American Revised*, are.

Additional community requests for WordNet are given by Clark et al. (2006): correlating event nominalizations with appropriate verb entries; relating verbs to their adjectival end-state descriptions, e.g., *store/stored*; reduction in the average number of words senses; increase in coverage.

While there has been extensive work attempting to manipulating and extend WordNet, we will touch upon just three groups in particular.

2.1.1 Automatic WordNet Supplementation

Snow, Jurafsky, and Ng (2006) automatically increased the size of WordNet approximately five-fold, primarily in the area of proper nouns such as people and places. The following year the same authors report on their work on addressing the sense overload problem (Snow et al., 2007) via a method allowing for user adjustable collapsing of senses.

While we find the Snow et al's underlying model intriguing, our recent experiments based on its results have left us skeptical of its utility within a larger system. For example, within the version of WordNet supplemented with 400k+ entries, there are over $3,000$ terms claiming to be children of one of the synsets for *President of the United States*[4]. As this synset also

[3] For example, *gold* is the hypernym of *gold dust*, and has as its own hypernym *noble metal*. By transitivity, *gold dust* is therefore either an instance or a kind of noble metal, neither of which is true

[4] "One of" : one sense refers to *the person who holds the office of head of state of the United States government*, the other to *the office of the United States head of state.*

contains the term *Chief Executive*, it is not surprising that the majority of these supposed hypoynyms turn out to be, in reality, the names of corporate CEOs. A similar example can be made for *boxer*, when it stands for a breed of dog. Under this node are over 100 entries, all of which are people, one of them being *Shaquille O Neal*.

We came to notice this issue with *boxer* through the evaluation of results of an attribute extraction algorithm that had made use of this resource (covered in a later section). *Boxer* was one of fifty randomly sampled classes, with its top ten extracted attributes to be manually evaluated. Owing to the instances being people instead of canines, there were zero correct results extracted for this class. This example helps to stress the point that Knowledge Acquisition is not a task that benefits anyone directly; it is a task meant purely to enable other systems to progress. As such, precision should be at least as much a concern here as elsewhere[5].

2.1.2 Manual WordNet Supplementation

The Omega Ontology, from Philpot, Hovy, and Pantel (2005), is the result of semi-automatically merging WordNet with a number of other manually created lexical resources. The project aims to provide a broad coverage semantic resource that makes minimal commitment to representational style. This in order that it may be maximally adopted by other researchers. We have recently gained access to this resource and are evaluating its potential utility in our work. For example, Omega contains a large collection of names of geographical entities, organized taxonomically, that we would like to import directly. However useful it may turn out to be, Omega lacks much of the vocabulary as aimed for by Snow, Jurafsky, and Ng (2006), especially proper names.

2.1.3 WordNet Conversion

WordNet synsets are paired with a dictionary style, definitional gloss. For example, *politician : a leader engaged in civil administration.* Mihalcea and Moldovan (2001) automatically parsed these glosses, converting the syntax trees into a Hobbsian[6] logical form. This so-called *eXtended WordNet*

[5] It should be stressed that the 400k expansion of WN represents the largest, and therefore most likely the noisiest, of the result sets offered by Snow.

[6] After the computational semanticist Jerry Hobbs, see (Hobbs et al., 1988) for more details.

(XWN) was then used to support abductive style inference for Question Answering.

We are interested in performing a similar conversion of WordNet glosses into logical forms, where it remains to be seen whether XWN's, pre-existing Hobbsian logical forms may be of use in that conversion process. Our own representational formalism, Schubert's Episodic Logic (EL) (Schubert and Hwang, 2000), is richer than that of Hobbs to an extent that may require processing the original texts rather than using XWN as is. As a simple example, XWN treats non-lexicalized compound nouns and prenominal adjectival modification as purely subsective. From the gloss, *psychological feature : a feature of the mental life of a living organism*, XWN contains the logical form:

```
psychological_feature:NN(x1) -> feature:NN(x1) of:IN(x1, x2)
                                mental:JJ(x2) life:NN(x2)
                                of:IN(x2, x3) living:JJ(x3)
                                organism:NN(x3)}
```

This says that a psychological feature is a feature of some x such that x is both `living` and an `organism`, where the feature is both `mental` and `life`.

This representation has shown to be useful in simple question answering, where one may iteratively remove predicates a from sentence's logical form according to a prespecified abductive weighting mechanism in the process of proving the answer to a question. In this case a formalism that breaks sentence meaning into as flat a form as possible may be preferred, as predicates can then be more easily filtered out individually with minimal impact on the global structure.

To contrast, the following is the same sentence in Episodic Logic:

(∀x : [x ((attr psychological.adj) feature.n)]
        (The y : [[y ((attr mental.adj) life.n)] and
                  [y of (K ((attr living.adj) organism.n))]]
        [x (feature-of.n y)]]))

Where: *mental* is taken as a predicative adjective that needs to be mapped into a attributive reading (giving as a result a predicate modifier, applied to *life*) via *attr*; *living organism* refers to a generic kind (as resulting from the operator K, mapping predicates to kinds); and *feature-of* is a function mapping the y that is both *mental life* and of the kind *living organism*, to the predicate selecting those x that are the *feature* of such y.

Our own practical motivation in pursuing large collections of knowledge is to enable existing work in the areas of continuous planning and machine self awareness. These projects employ EL as a representational base, and therefore KBs we wish to construct should conform to this representation. It is unclear at this point whether XWN may be transformed into useable EL statements automatically, or if instead we should construct our own such semantically parsed WordNet independently.

2.2 Cyc

The Cyc knowledge base is the result decades of man years of knowledge engineering. Recently it has become possible to acquire the majority of its contents for research purposes, in the aptly named ResearchCyc. This resource has as a result become the de facto knowledge base for a number of AI researchers, such as in the recent work of Ken Forbus and colleagues towards *Learning by Reading*, as described by Birnbaum et al. (2005).

Matuszek et al. (2006) describes the Cyc KB as being broken into three primary levels; the upper, middle and lower ontologies. The upper ontology stores high level meta-knowledge concerning structural classes for organizing the lower levels, along with most of the KB's knowledge of mathematics. The middle ontology stores the sort of common-sense knowledge that we are interested in acquiring automatically. The lower ontology stores facts about particular instances, e.g., information about chemical reactions, or properties of elements. It is this lower ontology that has been targeted by Cycorp's recent work in supplementing the KB via automated extraction; names of politicians, dates and locations of events, etc.

In contrast to the authors of Omega, Cycorp has made strong commitments to their own specific style of knowledge representation. As discovered by a fellow student, Fabrizio Morbini[7], representational design decisions made by Cycorp have left Cyc in a state that makes it nontrivial to automatically import any large portion of the KB into our own knowledge store. While Morbini has had partial success in this importation, he also reports what seem to be irreconcilable differences in ontological commitments, as well as practical issues in trying to determine the semantics of Cyc predicates, which tend to non-compositional concatenations of natural language phrases. In addition, Cyc is not "full", especially in the area of conditional relationships between events.

[7]Personal communication.

Chapter 3

Related Work

The following gives a representative survey of existing work in knowledge extraction.

3.1 Hyponymy Extraction

(Hearst, 1992) was one of the pioneers in the use of template driven extraction.

An example from the author:

```
The bow lute, such as the Bamabara ndang, is plucked and
has an individual curved neck for each string.
```

As explained by Hearst, native speakers of English have no trouble understanding that the *"Bamabara ndang"* is a *"bow lute"*, despite never having heard of a bow lute directly. This is an example is of what WordNet calls the *hyponymy* relation, which holds when:

> "native speakers of English accept sentences constructed from such frames as *An x is a (kind of) y*"

Working directly from this definition, Hearst developed a system based on mining text with regular expressions, all roughly of the form *"such NP as {NP,} *..."*. In reference to a previously performed study comparing parsing and pattern based acquisition, Hearst made the claim that pattern based methods are to be preferred due to their efficiency and increased accuracy.

Hearst's experiments targeted both an encyclopedia and a portion of the New York Times. From 8.6M words of text coming from the encyclopedia, 7067 sentences loosely fit a single template, with 152 instances fully matching. From the New York Times, of 20M words of text, 3178 sentences matched the given template, with 46 instances fully matching.

She raised in her analysis a number of problems that are now commonly recognized. The first of these is underspecification, such as in the included example *hyponym("steatornis","species")*, where the extracted relation could be said to be "too general". Alternatively, relations are sometimes extracted that are too context or point-of-view specific, such as *hyponym("Washington","nationalist")* or *hyponym("aircraft","target")*.

Regarding Hearst's expressed view towards pattern based extraction, today it is not clear that efficiency considerations are as important as they were then, although the majority of the community still operates under this belief. To begin with, commonsense knowledge extraction is presumably an offline task, not being waited on by some other service or user, so where is the rush? In addition, and especially given that computers are increasingly more powerful, we would need to see an example of a pattern based system tackling a collection of text that was then shown to be impractical to attempt with a parsing based system. Until that time, we maintain the belief that parsing based systems allow for a more natural encoding of the linguistic intuitions that go into an extractor. Note that these intuitions, pattern based or otherwise, need not be specified manually. As we point out later in this section, automatic acquisition of extraction patterns themselves, so called *seed-based* methods, are fast becoming the norm.

As to the second point, on the relative merits with respect to accuracy, we benefit from the fact that automatic parsing of open domain text has come a long way since 1992, as seen in such works as (Charniak, 2000; Collins, 1999; Lin, 2003; Hockenmaier, 2003). With that said, we do not preclude the use of regular patterns in our work, if it happens to be the case that the parsing tools at hand can be shown to repeatedly fail on a specific phenomenon of interest.

Moving away from direct reply to Hearst, the more important issue facing pattern based, vs parser based, extraction, is how to acquire knowledge of as rich a complexity as allowed for in natural language. Episodic Logic was designed specifically because of the wide ranging subtleties of meaning allowed for in human discourse, many times expressed through syntactic structure not easily captured by a small set of general purpose, regular patterns. It is this sort of complex knowledge we have in mind for the thesis and beyond. For example: A BIRD MAY BUILD A NEST IN WHICH TO LAY AND HATCH EGGS,[1] or IF A CAR COLLIDES WITH A PHYSICAL

[1] Due to Lenhart Schubert, personal communication.

OBJECT THEN THE DRIVER OF THAT CAR MAY BE INJURED AS A RESULT.

3.2 Learning QA Patterns

When annual Question Answering (Voorhees and Tice, 2000) competitions were first introduced, pattern based systems were not only the norm, they were also the best performing. Although some employed more sophisticated methods, most notably the efforts of Harabagiu, Pasca, and Maiorano (2000) and Harabagiu et al. (2000)[2], the best results came from those willing to simply sit down and author large collections of regular expressions (Soubbotin and Soubbotin, 2001).

Ravichandran and Hovy (2002) gave a process for deriving such answer extraction patterns automatically, serving as an exemplar for the many example based systems that have come since.

Ravichandran and Hovy acquired QA patterns in two stages: discover a set of patterns; calculate some sort of *a priori* accuracy weighting for each learned pattern.

Pattern discovery was performed as follows:

1. Pick a canonical example for the question type under consideration (such as BIRTHDATE or LOCATION).

2. Manually identify the primary question and answer terms. In their running example, *"When was Mozart born?"*, these terms become *"Mozart"* and *"1756"*.

3. Use these to query the web, keeping the top 1,000 documents. Automatically segment the sentences in each document.

4. Throw away all sentences that do not contain both the question and answer terms. Manually expand the terms into equivalence sets, e.g., {*"Mozart","Wolfgang Amadeus Mozart", "Mozart, Wolfgang Amadeus"*, *...*}. For any set of contiguous tokens in the sentence that match a member of an equivalence set, rewrite those tokens as the representative member.

[2] These more sophisticated systems eventually went on to show dominant performance; viewed by many as a welcome boost for "classical" AI technologies, and possibly helping to motivate the recent surge in interest in computational semantics.

5. Across all remaining sentences of all collected documents, find all contexts (substrings that lie between the question term and the answer term). Record respective counts for the occurrence of each unique context.

6. For each context, replace the terms with their canonical label, e.g. <NAME>, <ANSWER>, resulting in a candidate answer pattern for the given question.

With a set of generalized, frequency tagged patterns in place, one can then calculate a measure on the expected precision for each pattern:

1. Use the canonical question term to query the web, keeping the top 1,000 documents. Segment the sentences.

2. Delete all sentences that do not contain question terms.

3. For each remaining sentence, for each pattern, check for its presence in the sentence both in the case where it instantiates the known answer, and for other "noise" instantiations. Record these cases separately.

4. Delete those patterns that do not generate more than a set threshold of instances, e.g., "at least 5".

5. The precision of a pattern is then defined as the number of correct instantiations of a pattern over the total number of instantiations.

This procedure was applied to a handful of standard TREC question types, resulting in a large collection of extracted patterns. Performing QA testing on both the web and on the TREC QA corpus, they showed comparable performance to other systems at the time, with no effort placed towards manual pattern definition. Due to their web based accuracy being significantly higher than when restricted to the TREC corpus, the authors conclude that the web, due to greater redundancy of information, allows for wide recall using relatively simple patterns.

This last point is a common theme in modern systems: don't worry overly much about the coverage of your extraction patterns; assume your initial patterns will give acceptable levels of recall when applied to the entirety of the web, then focus on methods of post-processing results in order to determine accuracy (using procedures reliant on aggregate frequency). This process is in contrast to focusing on better language processing on the front end, in order to minimize extraction errors from the beginning. To some extent both strategies are needed; proper language processing requires background knowledge, just as acquiring background knowledge (we maintain) requires proper language processing. We assume an iterative process is required, going back and forth between acquiring knowledge at the limits of current processing, and then using that knowledge to enable the construction of richer extraction systems. Our difference with the mainstream of extraction research is that we feel that there already exists the capability to extract richer information than is currently being targeted by most, and that too much effort is being placed on developing measures exploiting redundancy of simple results as compared to pushing the state of the art in "old fashioned" linguistic analysis.

3.3 Paraphrase Rules

Lin and Pantel (2001a) extracted paraphrase rules using co-occurrence vectors based on syntactic paths. For example, the conjoining context between X and Y in sentences 1 and 2 are deemed similar if they tend to occur with the same terms as arguments.

(1) X finds a solution to Y.

(2) X solves Y.

Consider sentences 3 through 6.

(3) The commission found a solution to the strike.

(4) The committee found a solution to the crisis.

(5) The committee solved the problem.

(6) The government solved the problem.

The words *"commission"*, *"government"* and *"committee"* have high semantic affinity based on distributional similarity, as do *"strike"*, *"problem"* and *"crisis"*. Therefore, the conjoining contexts are deemed likewise similar. The authors then go one step further and rephrase "is similar to" as "implies". That is, *"The commission found a solution to the strike"* is considered to *imply* that *"The commission solved the strike."* is also valid; one is a semantics preserving paraphrase of the other. As they are considered mutual paraphrases, then they are imagined as being connected via biconditional, with empty slots as the X and Y arguments.

A high level description of their method:

1. Accumulate a large collection of sentences with accompanying dependency parses, in this case based on Lin's MINIPAR parser (Lin, 1998).

2. Identify the basic noun phrases within each sentence.

3. Collect all paths that connect similar nouns.

For example, the following:

```
N:subj:V<-buy->V:from:N
```

is a path extracted from sentence 7, containing a SUBJ link, and a FROM link, both extending from the verb *buy*.

(7) They had previously bought bighorn sheep from Comstock.

The authors report the following as the twenty most similar paths to (2):

```
Y is solved by X          Y is resolved in X
X resolves Y              Y is solved through X
X finds a solution to Y   X rectifies Y
X tries to solve Y        X copes with Y
X deals with Y            X overcomes Y
Y is resolved by X        X eases Y
X addresses Y             X tackles Y
X seeks a solution to Y   X alleviates Y
X do something about Y    X corrects Y
X solution to Y           X is a solution to Y
```

Recently Pantel, along with Ed Hovy and students have begun refining the results of the DIRT system by assigning semantic types to the X and Y arguments, as well as enforcing directionality on the paraphrase implication.

(Pantel et al., 2007) went back to the data from their work in 2001 to discover which sorts of arguments tended to fall in different positions of their paraphrases. They refer to their results as *Inferential Selectional Preferences (ISPs)*. In figure 3.1 we see the results for a rule equating learning and teaching.

The most likely arguments for X and Y are given in rank order, as according to their strength in similarity to the terms most commonly found in the given slots. Argument types, such as {*N370 culture, tradition, history*}, come from the results of Pantel and Lin's Clustering By Committee (CBC)

```
N:about:V<learn>V:from:N        N:about:V<teach>V:obj:N

X:
        {N370 culture, tradition, history}      19
        {N403 risk, danger, consequence}        9.5
        {N608 system, technology, process}      6.25
        {N94 issue, matter, question}           6.25
        {N392 China, Russia, Japan}             5.5
        {N1482 environment, lifestyle, life}    4.5
        {N350 strategy, principle, objective}   4.25
        {N276 courage, perseverance, hard work} 3.5
        {N174 action, activity, effort}         3.5
        {N259 magnitude, extent, scope}         3.25
Y:
        {N1565 him, them, me}                   75.25
        {N320 woman, child, man}                42
        {N702 people, citizen, resident}        14
        {N841 Anyone, everybody, everyone}      8.5
        {N623 Some, hundreds, thousands}        4
        {N188 revenue, profit, income}          2.5
```

Figure 3.1: Sample from the ISP ruleset, using CBC clusters

clustering system; a program for partitioning organizing words into distinct semantic clusters based on distributional similarity (Lin and Pantel, 2002).

The above rule might be interpreted as saying that, if you have a learning situation where the learning is *from*, e.g., a *woman*, and the learning situation is *about*, e.g., *culture*, then you may have a teaching situation where the object of verb is, e.g., a *woman* and the situation is *about*, e.g., *culture*.

Neither in their results, nor in personal communication do the authors commit to anything more or less than is expressed as in the figure; they are viewed exactly as syntactic paths that tend to share arguments.

We are interested in trying to derive something stronger, and are currently in the stages of transforming a subset of these results into Episodic Logic. Our target is the portion of rules connecting verbs that fall in the set of *happens-before* pairs, as according to the VerbOcean project (Chklovski and Pantel, 2004)[3].

Our first concern is the relative makeup of the DIRT ruleset. Since the extraction method requires that both X and Y are filled by the same, or *distributionally similar*, arguments, then the majority of the DIRT rules are very likely to be paraphrases (by default you expect any two given entities to have at most a single connecting relationship).

For example, George Bush is the president of the United States. He is also describable as the leader, the commander and chief, and the "face" of the United States[4]. Under loose constraints, these relationships may be considered paraphrases.

Now consider sentences 8 and 9.

(8) Jack searched for Mary.

(9) Jack found Mary.

Here we have shared arguments, but the events occur at different times. It is these sorts of relationships we are interested in here, and therefore like to recognize and extract from the ISP data. Table 3.2 contains statistics we've gathered on what we call the *happens-before* ISP subset. Figure 3.1 contains examples from our current transformation system.

The disparity between *hb.subj.obj.subj.obj* and *hb.subj.obj.subj.obj'* is explained by the fact that many of the ISP rules claim to have two terms per

[3]VerbOcean is described in section 3.4.

[4]He is also involved in other relationships with the US, such as being a citizen-of, but as the methods described rely on frequency, then it is only the frequently mentioned relationships that will be recognized.

ISP	12,377,742	*ISP.subj.obj.subj.obj*	1,629,144
hb	137,913	*hb.subj.obj.subj.obj*	31,076
hb.subj	113,628	*hb.subj.obj.subj.obj'*	4,312
hb.subj.subj	91,804		

Table 3.1: All values refer to counts. *ISP* is the entire set, *hb* is the subset of *ISP* containing a *happens-before* pattern. Suffixes refer to subsets with, e.g., at least one SUBJ node in one of the two relations (for SUBJECT).

```
(G e [(∃ Y  : [Y N195.bank.corporation.financial_institution]
             (∃ X  : [X N231.consumer.investor.customer]
                    (∃ Z  : [Z ((about X) detail)]
                           [Y offer Z]))) ** e]
     (∃ e' : [e' after e]
             [(∃ Z  : [Z ((about X) detail)]
                     [Y provide Z]) ** e']))
```

```
(G e [(∃ Y  : [Y N816.it.this.which]
             (∃ X  : [X N320.woman.child.man]
                    (∃ Z  : [Z ((about X) complaint)]
                           [Y receive Z]))) ** e]
     (∃ e' : [e' after e]
             [(∃ Z  : [Z ((about X) information)]
                     [Y provide Z]) ** e']))
```

```
(G e [(∃ Y (∃ X   : [X N604.Iran.Pakistan.Egypt]
                  (∃ Z  : [Z ((against X) attack)]
                         [Y stop Z]))) ** e]
     (∃ e' : [e' after e]
             [(∃ Z  : [Z ((on X) attack)]
                     [Y condemn Z]) ** e']))
```

Figure 3.2: Examples of our current output for mapping a subset of the ISP rules into Episodic Logic.

relation when in fact there are three, where the OBJ node was considered part of the larger path, and a free variable is placed within a verb modifier.

For example, the following DIRT rule,

```
N:about:N<concern<N:obj:V<address>V:subj:N
 <=>
N:of:N<issue<N:obj:V<resolve>V:subj:N
```

can be translated as:

```
"X addresses a concern about Y"
 <=>
"X resolves an issue of Y"
```

where *concern* and *issue* are treated only as part of the context.

Under discussion are such issues as the proper way to deal with modification when it occurs, such as "information about X", and whether it is possible to recognize when third arguments are shared across rules. For example, in figure 3.1 we don't want to say that a *complaint about X* is the same thing as *information about X*, but at the same time, if one were to *offer* you *details about X*, those are probably the same details as when described as you having been *provided* them. Whether third arguments are shared across verbs corresponds to the level of scoping we assume for the existential on Z; does it, or does it not, span the conditional? (We note that the use of a classic conditional here is certainly not right; most or all of the relationships here are only *possible* things that could happen, not guaranteed implications. Unfortunately the relative strength of the implication across ISP rules is not something we're likely to determine automatically.)

Finally, we note that the methods of Lin and Pantel have additional constraints that limit their usefulness for our purposes. We've mentioned their strict reliance on binary relations, but perhaps more important is that these X, Y arguments are required to be equivalent in the antecedent and consequent. It is not clear how their method could work without this constraint, and yet, sentence pairs such as (10) and (11) hint at how large a space of relations are being missed by the DIRT system.

(10) Jack said goodbye to Mary.

(11) Mary then drove to work.

3.4 Event Relations

VerbOcean (Chklovski and Pantel, 2004) is a an extracted collection of verb pairs assigned with one of a handful of relations, such as the earlier mentioned *happens-before*, as well as *similarity,strength,enablement,* and *antonymy.* Pairs were extracted and given weights as a function of frequency with which they matched patterns such as, in the case of *happens-before*, *"to X and subsequently Y"*. As discussed earlier, we make use of these pairs in our work as an initial filter on potential event/event relationships.

As a component of her doctoral thesis, Girju (2002) gave an extensive survey on lexical causality, surveying the many complex ways causality can be expressed in English. Her work in causal *extraction*, however, was focused specifically on discovering causal tendencies between event nominalizations, based purely on verb centered, infix contexts. For example, having automatically discovered that the conjoining context *set in motion* tends to denote causality, *(earthquake,tidal wave)* would be extracted based on a sentence such as (12).

(12) The earthquake set in motion a tidal wave.

Girju's emphasis was on a technique for automatically learning such contextual patterns, not on capturing the breadth of causal constructions that she had enumerated. Rather than going beyond *X pattern Y* extraction, she instead used the same approach to build an extractor to recognize *meronomy*, acquiring knowledge such as *"Helicopters have as parts rotors, windows, propellers, ..."*.

3.5 Large Scale Fact Extraction

Oren Etzioni leads a group well known for their work in large volume fact extraction, such as the systems KnowItAll(Etzioni et al., 2004), KnowIt-Now(Cafarella et al., 2005), and their most recent, TextRunner(Banko et al., 2007). Until recently these systems had been focused purely on the extraction of facts. With their most recent work they have begun exploring what they refer to as a new paradigm of extraction, *Open IE*, which involves a greater focus towards abstract assertions.

3.6 KNEXT

The work most relevant to our proposal is that done previously within our own research group. Schubert's *KN*owledge *EX*traction from *T*ext (KNEXT) program (Schubert and Tong, 2003; Schubert, 2002) has focused on the use of language *understanding* in the extraction process. Rather than look for instances of predefined relations, KNEXT is based around the idea of moving from sentence to sentence of an input document, trying to parse, transform everything it sees.

For example, consider the following glossed propositions from Schubert and Tong (2003):

> *"A US-STATE MAY HAVE HIGH SCHOOLS"*
>
> "CHILDREN MAY LIVE WITH RELATIVES"
>
> "A BOOK MAY BE WRITE-ED BY AN AGENT"

Logical form is created via a multi-stage process, each producing a conceptually distinct layer of logical form. As the KNEXT system is the primary component of our proposed infrastructure, we will now give an overview of the extraction process by considering the steps involved in processing sentence 13.

(13) The man kissed a woman.

First, an analysis of the syntax and surface level semantics gives the *initial logical form (ILF)*:

[⟨The man.n⟩ (⟨past kiss.v⟩ ⟨A woman.n⟩)]

Square brackets have for years been used in Episodic Logic to represent infix notation, where the subject argument is placed ahead of the predicate. The above logical form is isomorphic to:

((⟨past kiss.v⟩ ⟨A woman.n⟩) ⟨The man.n⟩)

Lexical items such as *man.n* or *kiss.v* are written with disambiguating part of speech information expressed in dot-suffix notation.[5] Further information may be provided, such as sense numbers, e.g., *man.n.1*.

[5] Such lexical annotation has historically been written as contained within trailing square brackets, such as *man[n]*. We have recently switched to the dot-suffix notation to avoid confusion from overuse of the square bracket, and to be more in line with the conventions of the community.

Continuing with the analysis, deciding on quantifier and tense operator scoping, and the introduction of variables, gives the *scoped logical form (SLF).*

```
(past (The x : [x man.n]
                (A y : [y woman.n]
                       [x kiss.v y])))
```

Finally, episodic variables are introduced as part of temporal deindexing, leaving *deindexed logical form (DLF).*

```
(∃ e : [e before NOW]
       [(The x : [x man.n]
                 (A y : [y woman.n]
                        [x kiss.v y])) ** e])
```

Language is meaningful only if there is a speaker (or writer) and a hearer (or reader). Language meaning is often dependent on one or both of these players, and as such are taken into account in the logical form:

```
(∃ u : [u same-time NOW]
       [[Speaker tell.v Hearer
         (that (∃ e : [e before NOW]
                      [(The x : [x man.n]
                                (A y : [y woman.n]
                                       [x kiss.v y])) ** e]))] ** u])
```

Not seen in the above example is the process of argument abstraction that is done during, or immediately after, the construction of the ILF. For example, if the above sentence were instead *Jack kissed a woman*, KNEXT would recognize *Jack* as a name commonly used to refer to a MALE-INDIVIDUAL, and would use this class category in place of *Jack* in logical form construction. This emphasizes that KNEXT is not concerned with the extraction of specific facts of particular events, but rather, aims to acquire the "ground rules" (implicitly) known by humans, and made use of in discourse.

While not directly motivated by such a description, this procedure happens to follow in a manner similar to what Carston (1988) gives as the required steps of a Gricean analysis, leading up to the pragmatic stage of handling *conversational implicature.* Differences include the stage at which ambiguous lexical terms are resolved, along with the handling of quantifier scope. Carston considers these, along with other steps, as part of distinct stage of understanding which he refers to as *conversational explicature*, which precedes understanding implicatures. We mention this as it touches on the limits of what we might be accomplished through the exploration of ideas discussed herein. Later we will discuss patterns relying on pragmatic assumptions, based around words such as *but* or *however*, whose analysis would need to take place in this currently non-existent layer dedicated to conversational implicature.

3.7 Summary

Much of the related work in extraction has been on static properties. For example, creating or extending ontologies, e.g., *If X is a dog then X is a canine* (Snow, Jurafsky, and Ng, 2006), or learning part-whole relations, e.g., *If X is a helicopter then X has a rotor* (Girju, Badulescu, and Moldovan, 2003).

Existing research in acquiring conditional relationships between events has tended to fall either in the area of paraphrase extraction, such as the work of Lin and Pantel (2001b) and Pantel et al. (2007), or between pairs of single words, be they verbs as in VerbOcean (Chklovski and Pantel, 2004) or nominalized events as in the system from Girju and Moldovan (2002).

Recent work such as that done by Patrick Pantel and his colleagues shows steady progress towards more complex sorts of knowledge. As impressive as this work is, it has yet to result in anything usable for complex inference; precision has been deferred in the name of recall, issues of knowledge representation and natural language semantics are almost entirely ignored. As will be seen in our section on Conditional Knowledge, we aim to approach the problem of knowledge extraction from a different position, one in line with the described KNEXT system.

Chapter 4

Attributes of Concept Classes

The following describes our efforts on the task attribute extraction for concept classes.

4.1 Initial Investigation

Contrasting prior work in data driven extraction, which had been focused on the extraction of predefined relations, Paşca and Van Durme (Paşca and Van Durme, 2007) built a data driven system for the acquisition of arbitrary properties, or *attributes* for a concept class. For example, instances of the class *Country* may have a *Capital City*, or instances of the class *Painter* may have the attribute *Works*.

Common nouns stand for sets of entities which share some characterizing set of attributes. Therefore, by searching for the attributes ascribed to individual members of a specified class, it should be possible to determine those attributes that are most commonly shared. This is especially true when those instances are named entities such as *France* or *President Carter*, which tend to unambiguously refer.

Novel to our work, we choose as our dataset a collection of search engine query logs, as compared to a corpus based on documents. Our hypothesis could be phrased as *"The attributes that define a class tend to be what people ask questions about."*

Our approach was as follows:

1. Specify target classes, e.g., *(Linux, MacOS, Solaris, ...)*, for the class

Class	Size	Examples of Instances
Drug	346	*Ibuprofen, Tobradex, Prilosec*
Company	738	*Ford, Xerox, Longs Drug Stores*
Painter	1011	*Georgia O'Keeffe, Ossip Zadkine*
City	591	*Hyderabad, Albuquerque, Tokyo*
Country	265	*Yemen, India, Paraguay, Egypt*

Table 4.1: Classes used, their respective number of instances, and examples.

OperatingSystem.

2. Collect candidate attributes through domain-independent patterns, e.g., *X of Y*. For example, the query string *"kernel size of Linux"* gives *Kernel Size* for class instance *Linux*.

3. Rank candidates based on weighted aggregate frequency, where full fledged natural language queries such as *"what is the population of iquitos"* have higher contribution than patterns manually determined *a priori* to be less promising.

4. Heuristically filter candidate list; discard candidate if part of a known proper name, e.g., *(Toyota) of Boston*, discard attributes subsumed by one of higher rank, where subsumption was determined by substring matching[1], e.g., *Street Map* vs. *Map*.

Our experiments were limited to a set of five classes, as seen in table 4.1. The classes *Drug, Company, Painter, City,* and *Country* correspond respectively to the general categories of artifacts, organizations, roles, and geographical entities.

We desired that discovered attributes would tend to not be shared by a large number of other concepts. For example, given our corpus, *Photo* is an attribute extracted for a large portion of instances, yet we did not believe it to be an attribute of much importance. This notion of importance was based on trying to find those attributes that were most *characteristic* of a given class. For example, *Studio* for *Painter* as compared to *Daughter*.

Note that decisions based on specified versus general importance do not necessarily align. Consider the attribute *Cost* for the class *Drug*; most would

[1]Which usually translates to an assumption of attributes being head final, with modifiers being subsective.

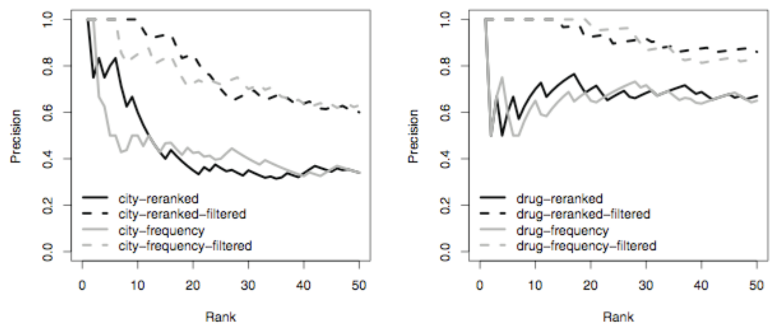

Figure 4.1: Precision as a function of rank. Dotted lines represent lists that have been heuristically filtered; black stands for lists reranked according to $Score_{reranked}$; grey represents $Score_{frequency}$.

consider cost an important attribute for a wide variety of classes, e.g., *Car*, *Food*, etc., and therefore a selection criterion based purely on specificity could remove attributes that humans presumably would not. We will return to this point later in our discussion of hierarchies.

Our original scoring function:

$$Score_{frequency}(C, A) = W_f(C, A)$$

Where W_f is the weighted aggregate frequency and C, A are a candidate class and attribute, respectively.

To filter non-class specific attributes we added a PMI-like factor (Turney, 2001) to the original formula:

$$Score_{reranked}(C, A) = W_f(C, A) \times S_f(C, A) \times \log \frac{W_f(C, A) \times N}{W_f(C) \times W_f(A)}$$

Where N is the total frequency over all class, attribute pairs and $S_f(C, A)$ is a smoothing factor to prevent emphasis of rare attributes. Precision results at various rank n for the classes *City* and *Drug* are presented in figure 4.1. We found that heuristic filtering had a dramatic effect on accuracy, with reranking by specificity having little effect.

Accuracy was determined based on labelling by hand the top 500 entries returned for each class in our initial experiments, then adding supplemental labels as needed if later revisions of the system brought previously unseen attributes into the top 50. Attribute candidates were labeled according to

Country: *capital, population, president, map, capital city, currency, climate, flag, culture, leader*
Drug: *side effects, cost, structure, benefits, mechanism of action, overdose, long term use, price, synthesis, pharmacology*
Company: *ceo, future, president, competitors, mission statement, owner, website, organizational structure, logo, market share*
City: *population, map, mayor, climate, location, geography, best, culture, capital, latitude*
Painter: *paintings, works, portrait, death, style, artwork, bibliography, bio, autobiography, childhood*

Table 4.2: Top ten attributes per class.

a simple, sliding scale; the labels *vital, OK, wrong* were given an accuracy weight of 1, 0.5 and 0. For example, *(Country, President)* would be *vital*, while *(City, Restaurant)* would merely be *OK*. Attributes such as *Users* for *Drug* or *Diary* for *Painter* were considered *wrong*.

We were encouraged by our results at rank ten, across all five classes, as seen in figure 4.2. Going to rank 50 we start to see a serious drop in precision. Especially in the case of *Company*, a number of errors at deeper rank come from the lesser senses of class instances. For example, *Volkswagen* may refer to the company, or the cars that the company produces (or the dealerships that sell the cars, etc.).

A glimpse of the difficulty inherent in determining what is and is not a *vital* versus *OK* attribute for a class can be seen in our discussion on class specificity. Yet more difficult is in determining recall of our method. We tackled this issue through a variety of comparisons.

Figure 4.3 lists the 13 most queried for attributes of *Country* in the first 1,893 questions used in the Question Answering track of TREC(Voorhees and Tice, 2000). As seen, six of the top ten $(1, 2, 10, 5, 3, 6)$ of our discovered attributes are direct string matches to the entries within the top ten TREC queries based on countries.

As another evaluation for the class *Country*, we compared our results against the attributes defined in the country tables of the CIA Factbook[2]. Unsurprisingly, we find that simple attributes such as *Flag* or *Map* are discovered, as compared to those such as *Household income consumption by percentage share* or *Manpower reaching military service are annually.* Figure 4.4 gives examples of this comparison.

[2] http://www.cia.gov/cia/publications/factbook

Attribute	Rank	Attribute	Rank
capital	1	*emperor*	-
population	2	*date of independence*	50
Prime Minister	19	*currency*	6
leader	10	*area*	23
capital city	5	*Queen*	76
President	3	*GDP*	32
size	20		

Table 4.3: Attributes from TREC questions ranked by frequency (top to bottom, left to right), with their corresponding positions in our rank list.

Attribute	Rank	Attribute	Rank
natural resources	21	*irrigated land*	-
terrain	132	*administrative divisions*	-
climate	7	*infant mortality rate*	-

Table 4.4: Examples of comparison against the CIA Factbook.

Attribute	R.	Attribute	R.
(*Painter, nationality*)	30	(*Painter, influences*)	11
(*Drug, is it addictive*)	-	(*Painter, awards*)	-
(*Drug, side effects*)	1	(*Country, income*)	88
(*Company,is it nonprofit*)	-	(*Country, neighbors*)	-
(*Company, competitors*)	4	(*City, mayor*)	3
(*City, quality of living*)	-	(*City, taxes*)	30

Table 4.5: Examples from user survey, matched against our results.

Finally, we performed a small survey in which participants were given two example classes with associated attributes[3], then asked to do the same for the five target classes. Users were not constrained to any minimum or maximum length. The results of this survey were most informative in that they pointed out our lack of attention to *unary* attributes, such as *Addictive?* for *Drug* or *Nonprofit?* for *Company*. The example results shown in figure 4.5 are based on conservatively equating the free form responses of participants to those in our result set.

Note the potential for formalizing these class, attribute pairs as formal, logical statements:

(∀x : [x country.n]
    (∃y : [y capital-of x]))

However, without additional work on refining results, the use of a universal as outside quantifier is premature. For example, we extracted the attribute *President* for the class *Country*, and yet there exist some countries for which this attribute does not hold. Instead these attributes need to be formalized as generic statements, whose interpretation allows for exceptions:

(G x : [x country.n]
    (∃y : [y president-of x]))

Alternatively, we would like to state such non-essential attributes paired with probabilities. However, it is unclear how these probabilities might be obtained, or how meaningful they might be. For instance, if we see

[3] *Dog* and *Planet*, with examples such as {*Size, Color, Breed*} and {*Composition, Distance from Earth, Moon(s)*}.

evidence of 20 countries having presidents, and assuming we know the full set of possible countries, that gives us a lower bound on the likelihood of an arbitrary country having a president, but we have no way of knowing whether the presidents of another 20 countries simply had no reason to be written about. This general problem of relating textual frequency to probabilities in the real world, while interesting and important to consider, is an issue well outside the bounds of this thesis.

4.2 Comparison of Data Sources

Recently we derived evidence giving support to our preference for query logs over documents (Paşca, Van Durme, and Garera, 2007). For an increased set of 20 classes[4], extraction results were compared between a system based on 50 million queries versus on run on top of 100 million web documents.

Both systems were based on the same manually defined query patterns. The one meaningful difference in extractors was that in the case of documents, more effort was needed in matching entities. Specifically, simple rules for determining membership within a larger NP were defined to limit faulty matches. As reported, the sentence *"Human activity has affected Earth's surface temperature during the last 130 years."* provides a valid match to the *HeavenlyBody* instance, *Earth*, while the sentence *"The market share of France Telecom for local traffic was 80.9% in December 2002."* does not provide a match for the *Country* instance *France*.

The accuracy of the query log based system was higher for all classes. At rank ten, the most dramatic differences came from such classes as *BasicFood* $(0.25, 1.0)$, or *HeavenlyBody* $(0.35, 1.0)$, with the given pairs referring to average precision for documents and queries, respectively.

It would appear that this difference in accuracy can be most strongly attributed to two factors: the primary sense of a term appears to dominate more strongly in queries than in web documents; our extraction patterns, such as *X's Y*, or *Y of X* less ambiguously denote attributes when used in a query.

For example, in reference to our first point, the incorrect attribute *Temple* extracted from documents for the class *HeavenlyBody* suggests a greater volume of document based text dedicated to Roman gods than in our query logs. This effect does not necessarily show queries to be better than docu-

[4] Actor, BasicFood, CarModel, CartoonCharacter, City, Company, Country, Drug, Flower, HeavenlyBody, Mountain, Movie, NationalPark, Painter, ProgLanguage, Religion, SoccerTeam, University, VideoGame, and Wine.

ments for all classes; a separate study may be beneficial to compare accuracy for classes with a high concentration of strongly ambiguous terms. And yet, these results do suggest that at least for whichever sense is dominant according to users of search engines, query logs are a more useful resource for attribute extraction, when relying on shallow extraction patterns.

As to our second point, the document derived attribute *Bowl* for the class *BasicFood* reinforces our hypothesis that query logs are specifically biased towards attributive uses of our extraction patterns.

This observed skew in the distribution of both pattern usage and term senses will be cited later in our discussions on corpus selection for targeted extraction of conditional relationships between episodes.

4.3 Generating Concept Classes

A requirement of our approach is prior definition of concept classes via the enumeration of instances. Paşca and Van Durme (2007) suggested that this issue should not be taken as a limitation because of the ease with which large numbers of concept classes could be gathered. However, recent investigations have revealed this to have been an overly optimistic assessment of the current state of available knowledge resources. For instance, we suggested one could simply amass classes based on existing lists, and gazetteers found online; this assumption about the abundance of manually generated data turned out to be false. There do exist the results previously discussed for automated mass supplementation of WordNet, but because our method's assumption of precisely defined input we were discouraged by the errors we found in sense selection (e.g., too many presidents).

Given the lack of pre-existing, high volume, high precision resources, we turned to another of our suggested options; using the results of a term clustering algorithm based on distributional similarity (Lin and Pantel, 2002). In this case we discovered that the granularity with which concepts are clustered using these frameworks does not necessarily correspond to that which we required for attribute extraction. For example, a cluster containing *George Bush* may include a number of other presidents, but it may also include other non-presidential politicians, their wives, foreign dignitaries, etc. Separately, even if a cluster was precise enough to contain only presidents, this would then mean that a hypothetical cluster of politicians must be bereft of the members of that class, as clusters are not arranged hierarchically.

Related to our consideration of WordNet based sources, we had access to the results of previous work in web scale hypernym extraction, as reported by (Pasca, 2004). Unfortunately these results on their own were also not of use, as the average precision was much too low.

As a result, we devised an approach for generating large numbers of concept classes by post processing the results of two separate systems into one: the first clustering noun phrases at a fixed level of granularity; the second generating large volumes of error prone hypernyms.

As discussed in (Paşca and Durme, 2008; Van Durme and Paşca, 2008), these resources can be combined using a measure akin to TFIDF to yield both high volume and high quality lists of instances representative of their class.

A summary of the method is as follows: while a given cluster of nouns may not be at the level of granularity one may desire, it still provides evidence on semantic relatedness. One might consider just those hypernym labels that are given to a reasonably large number of items in a cluster, and then further require that this label not occur too frequently for elements in other sets. If you then output just those elements with the given label, and only from the cluster under consideration, you end up a refined subset of your original hypernym data that tends to have higher average precision.

An obvious next step is to move towards the acquisition of corpus driven hierarchies, in order to be used in generalizing extraction results such as *Cost* for the classes *Drug*, *Car*, etc., to a higher level notion, e.g., *Artifact*.

4.4 Conclusions

We have performed a number of experiments investigating the task of extracting class attributes for concept classes. Our results suggest that for large numbers of classes we can provide at least ten high quality attributes that are characteristic for the given class.

Experiments in comparing query logs to webtext provides evidence that the accuracy of an extractor can be highly dependent on the underlying data. This is an intuitive result that should come to no surprise, yet research in extraction continues to be based on heterogeneous data crawled from the web.

Chapter 5

Acquiring Conditional Knowledge

Our goal is high precision commonsense knowledge, with a well developed semantics and usable for inference. Moving counter to tendencies within the field, we plan to focus on a formal understanding of the linguistic principles involved in the extraction process.

This is a large task that depends on a number of other areas, many of which are not yet fully understood. As such, the following is not a set of plans in order to "solve the conditional knowledge problem", however that might be formulated. Rather, these thoughts may be summarized as iteratively focusing on specific linguistic contexts, coming to a formal understanding of the syntactic and semantic phenomenon involved, then encoding this knowledge into an extraction system and evaluating its performance (whether automatically or otherwise). Previous efforts, such as Girju's, have presented a general awareness of many of the problems involved. However, extraction researchers continue to view binary relations centered on transitive verbs as cutting edge, state of the art.

In what follows we provide examples of contexts open for exploration, with a strong focus towards causality. Throughout our discussion a number of open problems are raised, such as proper determination of quantifier scoping and coreference resolution. A variety of methods may be considered to try and (at least partially) side step these issues. For instance, given the scale of data available online, one might develop techniques to *cherry pick* the most important sentences for a system to process, perhaps based on such simple heuristics as sentence length, or number of required prepositional phrase attachments (which are notoriously difficult to perform accurately on open domain text). Nevertheless, these outside problems are expected to remain as impediments for some time. We take the stance that these issues will be addressed in due course, but not by us. Our intention in what follows is to simply provide a road map for future work, highlighting issues to be tackled beyond what currently dominates the extraction literature.

5.1 Analytic (Periphrastic) Causatives

The most obvious sorts of sentences that convey causal information are those that directly include the word *cause*. Words such as *prevent*, or *make* may be defined directly in relation to the meaning of *cause*. This is in contrast to other words classically considered in causal literature, such as *kill*, *paint*, or *broke*. These later terms seem to convey a causal message, but with more detail provided as to either the causal event or the resultant state (the effect).

This first class of terms, *cause*, *make*, *etc.*, are known as either *analytic* or *periphrastic* causatives. The second class, *kill*, *paint*, *etc.*, are termed *lexical* causatives.

Jackendoff (1992) summarizes the distinction, leading with the following examples:

a. *Bill* { $^{?}$*broke the window* } *by startling the guy who was fixing it.*
caused the window to break
b. *Bill* { $^{?}$*killed Harry* } *on Tuesday by giving him poison on Monday.*
caused Harry to die
c. *The warm sun* { $^{?}$*grew the tomatoes.* }
made the tomatoes grow.

Jackendoff gives three reasons for the existence, and differing uses of lexicals as compared to periphrastics:

1. The *stereotypy effect*; if a lexical item exists, it tends to be used in preference to what would otherwise be a longer phrase based on a periphrastic.

2. Periphrastics allow for more complex modification. Compare *"Bill killed Harry on Monday* with *"Bill caused Harry to die on Monday"*. The modifier *"on Monday"* is ambiguous in the second sentence, allowing for temporal modification of only the resultant event, something

not possible when using the lexical. Likewise, Fodor (1970) contrasts *"John caused Bill to die on Sunday by stabbing him on Saturday"* against *"John killed Bill on Sunday by stabbing him on Saturday"*.

3. Idiosyncratic information may be encoded in the lexical. For example, sentence (c) shows that *"grew"* may select for an animate, purposeful subject, where-as periphrastics generally have no such restrictions.

We can take this to mean that lexical causatives exist to express causal relationships that have been identified as fitting generally recognized patterns. By making use of the knowledge implicitly stored in a lexical causative, a speaker can increase the average information content of his/her message, an assumed goal in human communication.

Use of a periphrastic causal may signal that the event is somehow either atypical or the speaker wishes to convey more precise information than would be given through use of the contextually appropriate lexeme.

Related to the second reason given above for the existence of these distinct classes, Goddard comments *"As we pass from analytic to morphological to lexical, we find increasing 'semantic cohesion', in some sense, between the causing event and the caused event..."* (Goddard, 1998). That is, the underlying events being described in a causal manner tend to be more closely bound when the speaker uses a lexical causative, especially with respect to temporal distance.

What does this mean for extraction? It means that we would prefer input based on analytic causatives, as less of their meaning is encapsulated in lexical semantics (which requires *a priori* knowledge of a lexeme's causal meaning), and is instead expressed in syntactic structure. It also means that we shouldn't expect them to occur as often as lexicals, since analytics are a tool to be used for constructing more detailed phrases usually only when a simple word like *kill* won't do.

We will return to the topic of lexical causatives in the next section.

Now we present an analysis of a simple causal sentence, using Episodic Logic, in order to give a flavor for the sorts of considerations we have in mind.

(14) An earthquake caused an emergency.

The initial logical form (ILF):

[⟨A earthquake.n⟩ ⟨past cause.v⟩ ⟨A emergency⟩]

After introduction of an event variable resulting from handling the 'past' operator, as well as determining scope for the noun phrases:

$$
\begin{array}{l}
(\exists\, \mathrm{e} : [\mathrm{e\ before\ Now}] \\
\quad [(\exists\, \mathrm{x} : [\mathrm{x\ earthquake.n}] \\
\qquad (\exists\, \mathrm{y} : [\mathrm{y\ emergency.n}] \\
\qquad\quad [\mathrm{x\ cause.v\ y}])) \ {**}\ \mathrm{e}])
\end{array}
$$

Note that both the subject and the object of sentence 14 are nominalized events, as compared to sentence 15.

(15) Jack caused an emergency.

Causal verbs have little restriction as to the semantic type of their subjects. However, we have for now decided that the cause-of.n relation will only take arguments that are events. Intuitively, when a causal verb takes a non-situational argument, the given argument is somehow *involved* in an event such that that event is either the cause or effect of the relationship being described. For example, when Jack causes an emergency, we take it that Jack performed some *action* [Jack|e][1], such that [Jack|e] resulted in a situation describable as an emergency.

We've recently begun to consider the use of the following rule, which will help unify the treatment of causal sentences[2]:

$$
\begin{array}{l}
(\forall\, \mathrm{x,y,e'} : [[\mathrm{x\ cause.v\ y}]\ {**}\ \mathrm{e'}] \\
\qquad [[\mathrm{e'\ extends\text{-}into\ y}] \wedge \\
\qquad\ (\exists\, \mathrm{p} : [\mathrm{p\ fluent\text{-}property}] \\
\qquad\qquad (\exists\, \mathrm{e} : [[\mathrm{e\ start\ e'}] \wedge [[\mathrm{x\ has\text{-}property\ p}]\ {**}\ \mathrm{e}]] \\
\qquad\qquad\quad [\mathrm{e\ cause\text{-}of.n\ y}]))])
\end{array}
$$

Where such fluent-properties are seen in sentences like: "An earthquake *occurred*", or "Jack *forgot to buy the plane tickets*". In the former, we might say that "An earthquake caused a tidal wave", which according to the suggested rule above would mean that there is some earthquake with a fluent property, which we might assume to be "occurrence", where the episode characterized by the earthquake occurring is the episode that was the cause-of the tidal wave.

[1] [Jack|e] is an example of an actor being paired with an event, the result being an action reference.

[2] Separately we say that if [e1 cause-of.n e2] then [(start-of e1) at-or-before (start-of e2)] and [(end-of e1) at-or-before (end-of e2)].

Determining what the fluent property is that a speaker meant when they give non-situational subjects, such as *Jack*, is a pragmatic issue, and not easily solved. In some cases the event description comes as a modifier, such as in sentence 16, or can potentially be retrieved from surrounding context, such as in sentence pair 17.

(16) Jack caused an emergency by forgetting to buy plane tickets.

(17) Jack caused an emergency. He forgot to buy plane tickets.

Continuing the analysis of sentence 14, we require the following axiom from Schubert (2000) :

$$(\exists)^{**} : \models \forall e.((\exists x\ \Phi)\ {**}\ e) \Leftrightarrow \exists x.(\Phi^{**}\ e)$$

This allows us to move existential quantifiers outside the scope of ** :

$$\begin{array}{l}
(\exists\, e\ :\ [e \text{ before Now}] \\
\qquad (\exists\, x\ [[[x \text{ earthquake.n}] \wedge \\
\qquad\qquad (\exists\, y\ :\ [y \text{ emergency.n}] \\
\qquad\qquad\qquad [x \text{ cause.v } y])]\ {**}\ e]))
\end{array}$$

Another axiom from Schubert (2000):

$$\begin{array}{l}
(\wedge)^{**} : \models \forall e.((\Phi \wedge \Psi)^{**} e) \Rightarrow \\
\qquad \exists e', e''.(\sqcup(e', e'') = e) \wedge (\Phi^{**}\ e') \wedge (\Psi^{**}\ e'')
\end{array}$$

which allows for:

$$\begin{array}{l}
(\exists\, e\ :\ [e \text{ before Now}] \\
\qquad (\exists\, x\ : \\
\qquad\qquad (\exists\, e', e''\ :\ (\sqcup(e', e'') = e) \\
\qquad\qquad\qquad [\,[[x \text{ earthquake.n}]\ {**}\ e'] \wedge \\
\qquad\qquad\qquad\ [(\exists\, y\ :\ [y \text{ emergency.n}] \\
\qquad\qquad\qquad\qquad [x \text{ cause.v } y])\ {**}\ e''\,]\,])))
\end{array}$$

earthquake.n is a what we call a *timeless* predicate: only events can be earthquakes, and intuitively an event either is or is not an earthquake; its "earthquakeness" is an atemporal property. Contrast this with a predicate such as student.n, which may be true for some interval, but in the worst case is limited to the lifespan of the person to which it is applied. True timeless predications characterize[3] the empty situation ϵ, and can roughly be thought of as "always true." If we let e' be the empty situation then ⊔(e',e") = e". Therefore e" = e.

(∃e : [e before Now]
    (∃x : [x earthquake.n]
        [(∃y : [y emergency.n]
            [x cause.v y]) ** e]))

Once more by (∃)** :

(∃e : [e before Now]
    (∃x : [x earthquake.n]
        (∃y : [[[y emergency.n] ∧[x cause.v y]] ** e])))

And once more by (∧)** :

(∃e : [e before Now]
    (∃x : [x earthquake.n]
        (∃y (∃e',e" : (⊔(e',e") = e)
            [[[y emergency.n] ** e'] ∧
            [[x cause.v y]] ** e"])))))

emergency.n is also a timeless predicate:

(∃e : [e before Now]
    (∃x : [x earthquake.n]
        (∃y : [y emergency.n]
            [[x cause.v y] ** e])))

Earthquakes and emergencies are events:

(∀x : [x earthquake.n] [x event.n])
(∀x : [x emergency.n] [x event.n])

Earlier we presented the following rule concerning fluent-properties:

(∀x,y : [x cause.v y]
    (∃p : [p fluent-property.n]
        (∃e : [[x has-property p] ** e]
            [e cause-of.n y])))

[3]That a sentence, S, may *characterize* an event, e, written [S ** e], is most the distinguishing feature of Episodic Logic.

When events are referenced directly as causes, the fluent property in question is always *occur*. That is, the occurrence of the earthquake is the property of the cause that can be specifically targeted as being the cause-of the result. As events are so often referenced directly as causes, it is useful to have a special case of the above rule:

$$
\begin{array}{l}
(\forall \mathrm{x,y,e'} : [[[\mathrm{x\ cause.v\ y}]\ {**}\ \mathrm{e'}] \wedge [\mathrm{x\ event.n}]] \\
\qquad\qquad [[\mathrm{e'\ extends\text{-}into\ y}] \wedge \\
\qquad\qquad\ (\exists \mathrm{e} : [[\mathrm{e\ start\ e'}] \wedge [[\mathrm{x\ occur.v}]\ {**}\ \mathrm{e}]] \\
\qquad\qquad\qquad\quad [\mathrm{e\ cause\text{-}of.n\ y}])])
\end{array}
$$

A separate meaning postulate states that if an event is characterized by something occurring, then that something is equivalent to the event so-characterized:

$$
\begin{array}{l}
(\forall \mathrm{e} : [\mathrm{e\ event}] \\
\qquad (\forall \mathrm{x} : [[\mathrm{x\ occur.v}]\ {**}\ \mathrm{e}] \\
\qquad\qquad \mathrm{x} = \mathrm{e}))
\end{array}
$$

Thus, from sentence 14 we know:

$$
\begin{array}{l}
(\exists \mathrm{e} : [\mathrm{e\ before\ Now}] \\
\qquad (\exists \mathrm{x} : [[\mathrm{x\ event}] \wedge [\mathrm{x\ earthquake.n}]] \\
\qquad\qquad (\exists \mathrm{y} : [[\mathrm{y\ event}] \wedge [\mathrm{y\ emergency.n}]] \\
\qquad\qquad\qquad [[\mathrm{x\ cause\text{-}of.n\ y}] \wedge [[\mathrm{x\ cause.v\ y}]\ {**}\ \mathrm{e}]])))
\end{array}
$$

Which can be simplified, as earthquakes and emergencies are events:

$$
\begin{array}{l}
(\exists \mathrm{e} : [\mathrm{e\ before\ Now}] \\
\qquad (\exists \mathrm{x} : [\mathrm{x\ earthquake.n}] \\
\qquad\qquad (\exists \mathrm{y} : [\mathrm{y\ emergency.n}] \\
\qquad\qquad\qquad [[\mathrm{x\ cause\text{-}of.n\ y}] \wedge [[\mathrm{x\ cause.v\ y}]\ {**}\ \mathrm{e}]])))
\end{array}
$$

Which can be glossed as *"An earthquake caused an emergency"*, as well as *"An earthquake was the cause of an emergency"*.

This may seem a long way to go for so little in return. However, the steps taken here are similar to those taken in more complicated arrangements. In addition, that our mechanics can bring us full circle is in fact an encouragement that we are on the right track.[4] Of primary importance

[4]Episodic Logic is motivated by the belief that language and mind are strongly connected; implying that a formal representation of knowledge should allow for seamless transition to and from language.

is that we were able to, in a formally justified manner, extract information given with respect to a particular episode into timeless predications that can be reasoned with more generally.

The next step in this analysis would be to compare the results for sentence 14 to the expected results from:

(18) An earthquake caused an emergency to worsen.

(19) An earthquake caused a building to collapse.

while at the same time keeping in mind the semantic oddness of sentences such as the following:

(20) An earthquake caused an emergency to collapse.

(21) An earthquake caused a building to worsen.

Note that these analyses are not given to show the steps we'd expect to be made by an applied extraction system. Rather, they are to be viewed as examples of a theory meant to guide the development of an applied system (i.e., to keep one honest). We assume extractors will make use of as many shortcuts and tricks as possible, while at the same time requiring our results to be representable, and deriveable, based on our original theory.

5.2 Lexical Causatives

A popular position taken wrt the semantics of lexical causatives (lexicals) such as *kill*, is that they are used as shorthand for something more explicit, such as *cause to die*. However, if we treat lexicals as just a paraphrase of a construction based on a core set of semantic primitives, then we run the risk of losing subtle aspects of meaning, such as seen earlier in the examples from Jackendoff. Consider the verbs *kill*, *murder*, and *assassinate*, which share a core meaning, but also have finer grain features that differentiate their usage. These finer features are difficult to robustly capture even for human lexicographers, let alone trying to acquire these distinctions automatically.

Thus, we commit ourselves to a weaker notion of paraphrasing that is still strong enough for our needs. Rather than saying that *kill* is the *equivalent* of *cause to die*, we instead say that *kill implies cause to die*. More specifically, using the terminology of EL, *cause to die* is a *partial description* of *kill*.

This is captured in the following meaning postulate for *kill*:

($\forall$ e,x,y : [[x kill.v y] ** e]

$$
\begin{aligned}
(\exists\, e' : \; & [[y \text{ die.v}] ** e'] \\
& [[x \text{ cause.v } e'] * e]))
\end{aligned}
$$

Thus, if *"Jack killed Bob."* characterizes the episode e, written loosely as:

[*"Jack killed Bob"* ** e]

Then *"Jack caused Bob to die."* is a partial description of e :

[*"Jack caused Bob to die"* * e]

A stronger form of this postulate adds to the consequent that e' should be a proper part of the event e, written as:

$$
\begin{aligned}
(\forall\, e,x,y : \; & [[x \text{ kill.v } y] ** e] \\
& (\exists\, e' : \; [[y \text{ die.v}] ** e'] \\
& \qquad [[e' \sqsubset e] \wedge [[x \text{ cause.v } e'] * e]]))
\end{aligned}
$$

This addition would be helpful in cases such as sentence 22, where intuitively it seems any temporal bounds placed on the killing event should also restrict the resultant death. Our modified MP would enable us to infer that *"Bob's death occurred yesterday"*, whereas our initial MP would not.

(22) Yesterday, Jack killed Bob.

Practically this means that if we had access to a lexicon mapping lexical causatives to a decomposition based on analytics, then even if there were subtle aspects of meaning being lost, we could at least use our tools built for analytics to get at the core causal component.

For such a lexicon we have in mind the detailed Appendix from the PhD thesis of Christoper Khoo (Khoo, 1995), which looked at whether recognition of causal relations could boost performance of an information retrieval system, based on a recognition system rooted in string level pattern matching. Within his dissertation is included a lengthy enumeration of lexical causatives, organized by category (such as "verbs of movement"), mapping base verbs to meanings rooted in analytics, typically *cause*.

5.3 Causal Adverbials

The theoretical approach introduced earlier becomes increasingly important as sentences become more complicated. The following is targeted at understanding causal adverbials, which to our knowledge have yet to be topic of an extraction system. Leading into the analysis we give a breakdown of sentence components, in order to highlight some of the aforementioned outside issues, as well the general importance of background knowledge in understanding even simple sentences.

(23) A car drove off the road, killing the driver.

5.3.1 *the road*

The noun phrase in *drove off the road* picks out the road implicitly given by a car driving. To know this is the case requires something such as the following script-like rule that assumes that if there is an event involving a car that is driving, then there is probably a road such that the car is on it:

$$
\begin{aligned}
(\mathrm{G\,e} : \ & (\exists \mathrm{x}\ [[[\mathrm{x\ car}] \wedge [\mathrm{x\ drive}]]\ {**}\ \mathrm{e}]) \\
& (\exists \mathrm{y}\ [[[\mathrm{y\ road}] \wedge [\mathrm{x\ on\ y}]]\ @\ \mathrm{e}]))
\end{aligned}
$$

The variable y is introduced through a process of accommodation, where it becomes available for resolution. G represents the generic quantifier[5], x is carried through as a dynamically bound variable. '@' is EL shorthand for denoting a given sentence characterizes an episode that is temporally concurrent with e. This formula would be glossed as *Cars drive on roads.*

5.3.2 *drove off the road*

Since the verb *drive* denotes some form of movement, the sentence may take a destination or direction specifying modifier, e.g., *to the store.* If we treat the prepositional phrase as this sort of modifier then we are asserting that at the beginning of the described event the car was not off the road, but was there by the end.

Alternatively the phrase may be a sentence modifier specifying that the activity of driving occurred, in its entirety, off the road.

[5]Generics operate like "soft" universals, e.g., *Birds fly.*. Knowing the strength of G is difficult, given the dependence on context; *Presidents have affairs, John walks to school, etc.*.

Between these, the most intuitive reading is that the phrase specifies direction and/or destination. We might infer this based on our knowledge that cars strongly tend to be driven on roads, but may sometimes leave them, usually by accident. Note that we might make the opposite conclusion if in the place of *A car* we found, e.g., *A Jeep*. This based on the knowledge that Jeeps are often thought of as *off road vehicles*.

5.3.3 *A car drove off the road*

Thus, we treat *A car drove off the road* to mean that a car was at some point in the past driving, presumably on the road, when at some later point it left that road.

5.3.4 *the driver*

We take the position that role specifying NPs are often used with a prepositional phrase gap that we'd gloss as *of Y*. That is, we treat the relational *the driver* as shorthand for *the driver of Y*. The implicitly given Y is introduced explicitly in the construction of the initial logical form, when we go from syntax to surface level semantics.

Inferring that Y is most likely the same entity as referred to by *the car* requires the knowledge that if X is driver of Y, then Y is typically some form of vehicle. A car is a type of vehicle, and as *a car* is the only such potential vehicle referent within this context, it is a safe assumption that Y is the car, and thus X is the driver of the car. Note that without such background knowledge we could just as easily infer that X is the driver of *the road*, or that X is simply "a driver", without any connection to the mentioned car.

Even in the presence of such background knowledge we might still make mistakes. It is not hard to create contexts where the above inference is no longer the only reasonable possibility. For example, consider if passage 24 were to precede 23.

(24) Yesterday evening a delivery truck driver pulled over to fix a flat tire. While attaching a spare, the man looked up to see a deer run into the heavy traffic.

5.3.5 LF

Our interest here is specifically on handling the adverbial, and thus give a step by step analysis of a simplified version of (23), which we give as sentence 25.

(25) A car crashed, killing John.

The initial logical form for 25:

[⟨a car⟩ ((adv-a (kill John)) ⟨past crash⟩)]

There are many possible scopings of this ILF. Without commenting on how we'd know which scopings to disregard, we will here just present the most intuitive reading. Beginning with the tense operator:

(∃ e : [e before Now]
  [[⟨a car.n⟩ ((adv-a (kill.v John)) crash.v)] ** e])

Introducing the existential:

(∃ e : [e before Now]
  [(∃ x : [x car.n]
    [x ((adv-a (kill.v John)) crash.v)]) ** e])

Action modifying adverbials are handled similarly as by Hwang and Schubert (1993), where rules were introduced for dealing with episode modifying adverbials such as *yesterday*.

(∃ e : [e before Now]
  [(∃ x : [x car.n]
    [[[x|e] kill.v John] ∧
     [x crash.v]]) ** e])

The following rule from Schubert (2000), which we used in our earlier analysis of earthquakes, allows us to lift the existential from outside the characterization relation, **.

(∃)** : ⊨∀e.((∃x. Φ) ** e) ⇔∃x.(Φ** e)

(∃ e : [e before Now]
  (∃x [[[x car.n] ∧
    [[[x|e] kill.v John] ∧
     [x crash.v]]] ** e]))

Atemporal sentences, which include episode modifying adverbials, may be lifted from outside the scope of the characterization relation, as we did earlier for *earthquake.n* and *emergency.n*.

(∃e : [e before Now]
    (∃x [[[x|e] kill.v John] ∧
        [[[x car.n] ∧[x crash.n]] ** e]]))

Earlier we gave a potential meaning postulate giving a connection between *kill* and *cause*. This MP required that the *killing* characterized some episode, but in the LF above we've claimed that the adverbial gave rise to an atemporal sentence. That is, we've given an analysis where the content of the adverbial does not characterize an event (other than the empty situation). Where did we go wrong?

The problem is that we relied on the machinery used for handling action-modifying adverbials such as *into a tree*. This example shows the need for a distinct treatment of causal adverbials within Episodic Logic, one allowing for the introduction of an additional event variable. Intuitively, we'd like sentences 26 and 27 to give the same LF.

(26) A car crashed, killing John.

(27) A car crashed. This killed John.

A corrected analysis of this example will require an extension of the theory of language associated with Episodic Logic.

5.4 Because

The following represents our current progress in detailing the distinction between *because* and *because of*.

In her thesis, Girju (2002) gives the sentence 28 as an example of a causal construction based on what she terms *prepositional causal links*. In sentence 29 we remove the relative clause stating the type of activity the book was involved in (ie *being carried by the man*). Sentence 30 is an attempt to get as close to the meaning of (28) by using *because* as a sentential connective, without adding information not given in sentence 28.

(28) A local man was kept off a recent flight because of a book he was carrying.

(29) A local man was kept off a recent flight because of a book.

(30) A local man was kept off a recent flight because he was carrying a book.

Notice that sentence 30 has a different causal focus than the previous sentences. There the causal event has to do specifically with *carrying*, as compared to some other property that the book possesses besides its involvement in the carry event. For example, contrast sentence 30 with 31.

(31) A local man was kept off a recent flight because a book he was carrying appeared suspicious.

Sentence 32 has a similar meaning to (30), through a nominalization of the carry event. Again, the meaning in this case is not the same as that expressed in the original sentence.

(32) A local man was kept off a recent flight because of his carrying of a book.

It seems that when using the prepositional connective *because of*, when the noun under causal focus is modified by an eventive relative clause, the event thus introduced is not the cause intended by the speaker.

This conclusion is further supported by common expressions such as sentence 33, where it is not simply the act of saying that is being identified as the cause, rather, we are concerned more with the content of what was said.

(33) It was because of something he said.

Skipping the intermediate steps, our proposed LF for sentence 28:

```
(∃x : [x ((attr local) man)]
      (∃z : [z ((attr recent) flight)]
            [(∃e' : [e' before Now]
                    [[x (pasv keep-off) z] ** e'])
            because (∃e : [e at-or-before e']
                          (∃y : [y book]
                                [[x (prog (carry y))] ** e]))]))
```

The contextual pattern *X because of Y* requires that X be a proposition, such as in sentences 34, 35, and 36, all of which come from the Brown corpus.

(34) He missed the 1955 season because of an operation on the ailing knee, then played 77 minutes in 1956.

(35) There is little enthusiasm for spending money to develop more powerful engines because of the erroneous belief that the aircraft has been made obsolete by the missile.

(36) He could not leave the road because of the water-filled drainage ditch.

Examples such as these, especially (34), leaves us optimistic as to the use of the *X because of Y* pattern. X appears to often be a declarative sentence simple enough to promise accurate parsing; a condition we can check for heuristically based on measures such as sentence length and depth. Y will tend to be a nominalized event, which as we've shown earlier may be directly mapped into a characterizing description of the relevant episodic variable, which in this case will be the cause.

5.5 Causative/Inchoative Alternation

The causative/inchoative alternation describes the relationship between lexical causatives which share form but differ in transitivity, such as *burned* (transitive) and *burned* (intransitive). The most important aspect of this alternation is that when the transitive version is used, what would have been subject in the intransitive case is now object, and the new subject is ascribed cause status; *The newspaper burned* as compared to *The fire burned the newspaper*.

Jack broke the vase may be understood as *Jack caused the vase to break*, or *The vase broke because of Jack*.

This relationship is also used to describe the connection between verbs and adjectives that share forms. To borrow an example from Goddard, contrast *"I cleaned the car"* with *"The car was clean"* (Goddard, 1998).

As the they share forms, the connection between verb and adjective is easy to make automatically. Using *clean* as an example, this alternation allows for a direct set of lexical axioms resembling something of the form:

(∃ e,x,y : [[x clean.v y] ** e]
    (∃ e',e" : [[y clean.a] ** e']
        [[e cause-of.n e'] ** e"]))

5.6 Pragmatic Contexts

While pragmatic understanding is generally considered more difficult to automate than semantics, there exist specific pragmatic implications that lie almost as close to the surface as the semantic phenomena we've discussed thus far.

In particular, we are interested in default inferences able to be made in the presence of negative contexts, such as evoked through the use of *but*, *despite*, or *and yet*. For example, from sentence 37 we may profitably assume *"If a man likes a woman, then NOT(the man will decide not to ask her out)"*, in other words, he *will* decide to ask her out.

(37) Jack liked Mary but decided not to ask her out.

According to Blakemore (2000), *but* carries with it little information other than serving as a conjunction that in some way *cancels out* an implication pragmatically derivable from the first of the items being conjoined. While Blakemore stresses the difficulty in determining just what precisely is being cancelled, our analysis of sentences from the Brown corpus suggests that one may quite regularly assume the effect of direct negation of the second of the propositions being conjoined.

The following additional examples come from the Brown corpus:

(38) Wexler admitted in earlier court hearings that he issued grand jury subpoenas to about 200 persons involved in the election investigation, questioned the individuals in the Criminal courts building, *but* did not take them before the grand jury.

(39) Commandeering a passing car, Kimmell pursued the fleeing vehicle, *but* lost it in traffic.

(40) Kern began reading a lot about the history and philosophy of Communism, *but* never felt there was anything he, as an individual, could do about it.

(41) *Despite* the lip service paid by local governments, the anti-secrecy statutes have been continuously subverted by reservations and rationalizations.

(42) The Eisenhower budget was simultaneously inadequate in its provisions *and yet* extravagant in its projections of revenue to be received.

Blackmore points out that *never-the-less* and *however* may be substituted for *but* in particular contexts, which suggests to us the possibility of additional constraints that might benefit the extraction process.

5.7 Generics in Encyclopedias

Simple generic statements can be used to convey a large portion of the background knowledge we would like to possess, for instance (43).

(43) Dogs bark.

Unfortunately sentences of this form do not often appear in corpora based on sources such as newswire text, where general properties are assumed to be understood, and the text is devoted to details on how the states of various entities have recently changed (that is, "what's new").

Encyclopedias are a prime source for these sorts of declarative statements, yet, like newswire, they tend to be difficult to accurately parse automatically. In this way they are more like general text but with a fixed topic, as compared to the short, more easily processed definitions in dictionaries, such as our earlier examples from WN.

Recently we had begun to explore the use of the Children's version of the Encyclopedia Britannica Online[6], in the hopes of finding sentences of average length and complexity less than what we observed in other data sources. Shortly after our initial investigation this resource was made no longer accessible online. This is unfortunate due to the ease in which we found examples as (44).

(44) City buses operate within cities and travel at slow speeds.

Alternatively one might try and make use of a very large encyclopedia, such as Wikipedia, and cherry pick sentences which suggest being accurately parsed and interpreted.

5.8 Summary

We've covered a series of phenomena and resources that might allow for the automatic acquisition of conditional relationships typically holding between events. If one were to take these ideas to the point of constructing an extraction system, we've a proposed methodology based on an iterative approach of first giving a formal analysis to a relevant linguistic phenomenon, building this knowledge into an existing extraction system, then collecting, evaluating results to guide the next round of development.

We've been clear about the fact that the required tools, such as high precision software for open domain syntactic parsing, may not yet be at the stage to support high volume extraction results with an average level precision that will be directly usable for inference. However, we believe that progress in these outside areas will continue to improve. In comparison, we are less optimistic about the current state of computational semantics with regards to knowledge acquisition, which continues to remain strongly data driven, even as it forces the use of impoverished knowledge representations.

[6] http://www.britannica.com/.

We offer these notes as an example of an alternative path towards building large scale, common sense knowledge bases from textual resources.

References

Banko, Michele, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, Hyderabad, India.

Birnbaum, L., K. Forbus, E. Wagner, J. Baker, and M. Witbrock. 2005. Analogy, intelligent ir, and knowledge integration for intelligence analysis. In *AAAI Spring Symposium 2005 Workshop on AI Technologies for Homeland Security*.

Blakemore, Diane. 2000. Indicators and procedures: nevertheless and but. In *Journal of Linguistics*, pages 463–486. Cambridge University Press.

Cafarella, Michael J., Doug Downey, Stephen Soderland, and Oren Etzioni. 2005. Knowitnow: Fast, scalable information extraction from the web. In *HLT-EMNLP*.

Carston, Robyn. 1988. Implicature, Explicature, and Truth-Theoretic Semantics. In Ruth M. Kempson, editor, *Mental Representations: The Interface Between Language and Reality*. Cambridge University Press, pages 155–181.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL)*, pages 132–139, Seattle, Washington.

Chklovski, Timothy and Patrick Pantel. 2004. In *VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations*, Barcelona, Spain. Conference on Empirical Methods in Natural Language Processing (EMNLP-04).

Clark, Peter, Phil Harrison, Tom Jenkins, John Thompson, and Rick Wojcik. 2006. From wordnet to a knowlege base. In *AAAI 06 Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*. AAAI.

Collins, Michael John. 1999. *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Etzioni, Oren, Michael Cafarella, Doug Downey, Stanley Kok, AnaMaria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2004. Web-scale information extraction in knowitall. In *WWW*.

Girju, R., A. Badulescu, and D. Moldovan. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Conference*, Edmonton, Canada.

Girju, R. and D. Moldovan. 2002. In *"Mining Answers for Causation Questions"*, Stanford University, California, March. American Association for Artificial Intelligence (AAAI) - Spring Symposium.

Girju, Roxana. 2002. *Text Mining for Semantic Relations*. Ph.D. thesis, The University of Texas at Dallas.

Goddard, Cliff. 1998. *Semantic Analysis : A Practical Introduction*. Oxford University Press, May.

Harabagiu, Sanda, Dan Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeam, Razvan Munescu, Roxana Girgu, Vasile Rus, and Paul Moraresc u. 2000. FALCON: Boosting knowledge for answer engines. In *Proceedings of the Text Retrieval Conference (TREC-9)*.

Harabagiu, Sanda, Marius Pasca, and Steven Maiorano. 2000. Experiments with open-domain textual question answering. In *Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000)*, Saarbrücken, Germany, August.

Hearst, Marti. 1992. "automatic acquisition of hyponyms from large text corpora". In *COLING*.

Hobbs, Jerry R., Mark Stickel, Paul Martin, and Douglas D. Edwards. 1988. Interpretation as abduction. In *26th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, pages 95–103, Buffalo, New York.

Hockenmaier, Julia. 2003. *Data and Models for Statistical Parsing with Combinatory Categorial Grammar*. Ph.D. thesis, University of Edinburgh.

Hwang, Chung Hee and Lenhart K. Schubert. 1993. Interpreting temporal adverbials. In *HLT '93: Proceedings of the workshop on Human Language Technology*, pages 138–143, Morristown, NJ, USA. Association for Computational Linguistics.

Jackendoff, Ray S. 1992. *Semantic Structures*. MIT Press.

Kaplan, Aaron N. and Lenhart K. Schubert. 2001. Measuring and improving the quality of world knowledge extracted from wordnet. Technical Report 751, University of Rochester, Computer Science Department, Rochester, New York 14627, May.

Khoo, Christopher S.G. 1995. *Automatic Identification of Causal Relations in Text and Their Use for Improving Precision in Information Retrieval*. Ph.D. thesis, Syracuse University.

Lin, Dekang. 1998. Dependency-based evaluation of minipar. In *Proceedings of the Workshop on Evaluation of Parsing Systems at LREC-98*, Granada, Spain.

Lin, Dekang. 2003. Dependency-based evaluation of minipar. In Ann Abeille, editor, *Building and using Parsed Corpora*, Language and Speech series. KLUWER, Dordrecht.

Lin, Dekang and Patrick Pantel. 2001a. Dirt - discovery of inference rules from text. In *ACM Conference on Knowledge Discovery and Data Mining (KDD-01)*, pages 323–328.

Lin, Dekang and Patrick Pantel. 2001b. "discovery of inference rules for question answering". *Natural Language Engineering*, 7(4):343–360.

Lin, Dekang and Patrick Pantel. 2002. Concept discovery from text. In *Proceedings of the 19th international conference on Computational linguistics (COLING-02)*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

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

Mihalcea, R. and D. Moldovan. 2001. extended Wordnet: Progress Report. In *NAACL 2001 - Workshop on WordNet and Other Lexical Resources*.

Miller, George A. and Florentina Hristea. 2006. Wordnet nouns: Classes and instances. *Computational Linguistics*, 32(1):1–3.

Paşca, Marius and Benjamin Van Durme. 2008. Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs. In *Proceedings of ACL*, Columbus, Ohio.

Paşca, Marius and Benjamin Van Durme. 2007. What you seek is what you get: Extraction of class attributes from query logs. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 2832–2837, Hyderabad, India.

Paşca, Marius, Benjamin Van Durme, and Nikesh Garera. 2007. The role of documents vs. queries in extracting class attributes from text. In *ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007)*, Lisboa, Portugal, November 6-9.

Pantel, Patrick, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. Isp: Learning inferential selectional preferences. In *North American Association for Computational Linguistics / Human Language Technology (NAACL HLT 07)*, pages 564–571, Rochester, NY, USA.

Pasca, Marius. 2004. Acquisition of categorized named entities for web search. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 137–145, New York, NY, USA. ACM Press.

Pedersen, Ted. 2007. Wordnet-similarity bibliography. `http://www.d.umn.edu/~tpederse/wnsim-bib/`.

Philpot, Andrew, Edward Hovy, and Patrick Pantel. 2005. The omega ontology. In *Proceeding of the ONTOLEX Workshop at the International Conference on Natural Language Processing (IJCNLP)*.

Ravichandran, Deepak and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the ACL Conference*.

Schubert, Lenhart K. 2000. The Situations We Talk About. *Logic-based Artificial Intelligence*, pages 407–439.

Schubert, Lenhart K. 2002. In *Can we derive general world knowledge from texts?*, pages 94–97, San Diego, CA, March 24-27. HLT.

Schubert, Lenhart K. and Matthew H. Tong. 2003. Extracting and evaluating general world knowledge from the brown corpus. In *Proceedings of the*

HLT/NAACL Workshop on Text Meaning, Edmonton, Alberta, Cananda, May 31.

Schubert, L.K. and C.H. Hwang. 2000. Natural language processing and knowledge representation: Language for knowledge and knowledge for language.

Snow, Rion, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING/ACL (2006)*.

Snow, Rion, Sushant Prakash, Daniel Jurafsky, and Andrew Y. Ng. 2007. Learning to merge word senses. In *Proceedings of EMNLP 2007*.

Soubbotin, M.M. and S.M. Soubbotin. 2001. Patterns of potential answer expressions as clues to the right answer. In *Proceedings of the TREC-10 Conference*, page 175182, Gaithersburg, MD.

Turney, P. 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings ECML-2001*, Freiburg, Germany.

Van Durme, Benjamin, Yifen Huang, Anna Kupsc, and Eric Nyberg. 2003. Towards light semantic processing for question answering. In *Proceedings of HLT/NAACL (2003) Workshop on Text Meaning*, Edmonton, Cananda, May 31.

Van Durme, Benjamin and Marius Paşca. 2008. Finding Cars, Goddesses and Enzymes: Parametrizable Acquisition of Labeled Instances for Open-Domain Information Extraction. In *Proceedings of AAAI*, Chicago, Illinois.

Voorhees, E. and D. Tice. 2000. Building a question answering test collection. In *Proceedings of SIGIR-2000*, pages 200–207, July.