# Deriving Generalized Knowledge from Corpora using WordNet Abstraction

**Benjamin Van Durme**, **Phillip Michalak** and **Lenhart K. Schubert**
Department of Computer Science
University of Rochester
Rochester, NY 14627, USA

## Abstract

Existing work in the extraction of commonsense knowledge from text has been primarily restricted to factoids that serve as statements about what *may possibly* obtain in the world. We present an approach to deriving stronger, more general claims by abstracting over large sets of factoids. Our goal is to coalesce the observed nominals for a given predicate argument into a few predominant types, obtained as WordNet synsets. The results can be construed as generically quantified sentences restricting the semantic type of an argument position of a predicate.

## 1 Introduction

Our interest is ultimately in building systems with commonsense reasoning and language understanding abilities. As is widely appreciated, such systems will require large amounts of general world knowledge. Large text corpora are an attractive potential source of such knowledge. However, current natural language understanding (NLU) methods are not general and reliable enough to enable broad assimilation, in a formalized representation, of explicitly stated knowledge in encyclopedias or similar sources. As well, such sources typically do not cover the most obvious facts of the world, such as that ice cream may be delicious and may be coated with chocolate, or that children may play in parks.

Methods currently exist for extracting simple "factoids" like those about ice cream and children just mentioned (see in particular (Schubert, 2002; Schubert and Tong, 2003)), but these are quite weak as general claims, and – being unconditional – are unsuitable for inference chaining. Consider however the fact that when something is said, it is generally said by a person, organization or text source; this a conditional statement dealing with the potential agents of *saying*, and could enable useful inferences. For example, in the sentence, "*The tires were worn and they said I had to replace them*", *they* might be mistakenly identified with *the tires*, without the knowledge that *saying* is something done primarily by persons, organizations or text sources. Similarly, looking into the future one can imagine telling a household robot, "*The cat needs to drink something*", with the expectation that the robot will take into account that if a cat drinks something, it is usually water or milk (whereas people would often have broader options).

The work reported here is aimed at deriving generalizations of the latter sort from large sets of weaker propositions, by examining the hierarchical relations among sets of types that occur in the argument positions of verbal or other predicates. The generalizations we are aiming at are certainly not the only kinds derivable from text corpora (as the extensive literature on finding *isa*-relations, partonomic relations, paraphrase relations, etc. attests), but as just indicated they do seem potentially useful. Also, thanks to their grounding in factoids obtained by open knowledge extraction from large corpora, the propositions obtained are very broad in scope, unlike knowledge extracted in a more targeted way.

In the following we first briefly review the method developed by Schubert and collaborators to abstract factoids from text; we then outline our approach to obtaining strengthened propositions from such sets of factoids. We report positive results, while making only limited use of standard

*Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 808–816,
Athens, Greece, 30 March – 3 April 2009. ©2009 Association for Computational Linguistics

corpus statistics, concluding that future endeavors exploring knowledge extraction and WordNet should go beyond the heuristics employed in recent work.

## 2 KNEXT

Schubert (2002) presented an approach to acquiring general world knowledge from text corpora based on parsing sentences and mapping syntactic forms into logical forms (LFs), then gleaning simple propositional factoids from these LFs through abstraction. Logical forms were based on Episodic Logic (Schubert and Hwang, 2000), a formalism designed to accommodate in a straightforward way the semantic phenomena observed in all languages, such as predication, logical compounding, generalized quantification, modification and reification of predicates and propositions, and event reference. An example from Schubert and Tong (2003) of factoids obtained from a sentence in the Brown corpus by their KNEXT system is the following:

> *Rilly or Glendora had entered her room while she slept, bringing back her washed clothes.*

```
 A NAMED-ENTITY MAY ENTER A ROOM.
 A FEMALE-INDIVIDUAL MAY HAVE A ROOM.
 A FEMALE-INDIVIDUAL MAY SLEEP.
 A FEMALE-INDIVIDUAL MAY HAVE CLOTHES.
 CLOTHES CAN BE WASHED.

((:I (:Q DET NAMED-ENTITY) ENTER[V]
                (:Q THE ROOM[N]))
 (:I (:Q DET FEMALE-INDIVIDUAL) HAVE[V]
                (:Q DET ROOM[N]))
 (:I (:Q DET FEMALE-INDIVIDUAL) SLEEP[V])
 (:I (:Q DET FEMALE-INDIVIDUAL) HAVE[V]
                (:Q DET (:F PLUR CLOTHE[N])))
 (:I (:Q DET (:F PLUR CLOTHE[N])) WASHED[A]))
```

Here the upper-case sentences are automatically generated verbalizations of the abstracted LFs shown beneath them.[1]

The initial development of KNEXT was based on the hand-constructed parse trees in the Penn Treebank version of the Brown corpus, but subsequently Schubert and collaborators refined and extended the system to work with parse trees obtained with statistical parsers (e.g., that of Collins (1997) or Charniak (2000)) applied to larger corpora, such as the British National Corpus (BNC), a 100 million-word, mixed genre collection, along with Web corpora of comparable size (see work of Van Durme et al. (2008) and Van Durme and Schubert (2008) for details). The BNC yielded over 2 factoids per sentence on average, resulting in a total collection of several million. Human judging of the factoids indicates that about 2 out of 3 factoids are perceived as reasonable claims.

The goal in this work, with respect to the example given, would be to derive with the use of a large collection of KNEXT outputs, a general statement such as *If something may sleep, it is probably either an animal or a person.*

## 3 Resources

### 3.1 WordNet and Senses

While the community continues to make gains in the automatic construction of reliable, general ontologies, the WordNet sense hierarchy (Fellbaum, 1998) continues to be the resource of choice for many computational linguists requiring an ontology-like structure. In the work discussed here we explore the potential of WordNet as an underlying concept hierarchy on which to base generalization decisions.

The use of WordNet raises the challenge of dealing with multiple semantic concepts associated with the same word, i.e., employing WordNet requires *word sense disambiguation* in order to associate terms observed in text with concepts (*synsets*) within the hierarchy.

In their work on determining selectional preferences, both Resnik (1997) and Li and Abe (1998) relied on uniformly distributing observed frequencies for a given word across all its senses, an approach later followed by Pantel et al. (2007).[2] Others within the knowledge acquisition community have favored taking the first, most dominant sense of each word (e.g., see Suchanek et al. (2007) and Paşca (2008)).

As will be seen, our algorithm does not select word senses prior to generalizing them, but rather as a byproduct of the abstraction process. Moreover, it potentially selects multiple senses of a word deemed equally appropriate in a given context, and in that sense provides *coarse-grained* disambiguation. This also prevents exaggeration of the contribution of a term to the abstraction, as a result of being lexicalized in a particularly fine-grained way.

### 3.2 Propositional Templates

While the procedure given here is not tied to a particular formalism in representing semantic con-

[1] Keywords like `:i`, `:q`, and `:f` are used to indicate infix predication, unscoped quantification, and function application, but these details need not concern us here.

[2] Personal communication

text, in our experiments we make use of *propositional templates*, based on the verbalizations arising from KNEXT logical forms. Specifically, a proposition $F$ with $m$ argument positions generates $m$ templates, each with one of the arguments replaced by an empty *slot*. Hence, the statement, A MAN MAY GIVE A SPEECH, gives rise to two templates, A MAN MAY GIVE A __, and A __ MAY GIVE A SPEECH. Such templates match statements with identical structure except at the template's slots. Thus, the factoid A POLITICIAN MAY GIVE A SPEECH would match the second template. The slot-fillers from matching factoids (e.g., MAN and POLITICIAN form the input lemmas to our abstraction algorithm described below.

Additional templates are generated by further weakening predicate argument restrictions. Nouns in a template that have not been replaced by a free slot can be replaced with an *wild-card*, indicating that anything may fill its position. While slots accumulate their arguments, these do not, serving simply as relaxed interpretive constraints on the original proposition. For the running example we would have; A __ MAY GIVE A ?, and, A ? MAY GIVE A __, yielding observation sets pertaining to things that may give, and things that may be given.[3]

We have not restricted our focus to two-argument verbal predicates; examples such as A PERSON CAN BE HAPPY WITH A __, and, A __ CAN BE MAGICAL, can be seen in Section 5.

## 4 Deriving Types

Our method for type derivation assumes access to a word sense taxonomy, providing:

$\mathcal{W}$ : set of words, potentially multi-token
$\mathcal{N}$ : set of nodes, e.g., word senses, or *synsets*
$\mathcal{P} : \mathcal{N} \rightarrow \{\mathcal{N}^*\}$ : parent function
$\mathcal{S} : \mathcal{W} \rightarrow (\mathcal{N}^+)$ : sense function
$\mathcal{L} : \mathcal{N} \times \mathcal{N} \rightarrow \mathbb{Q}^{\geq 0}$ : path length function

$\mathcal{L}$ is a distance function based on $\mathcal{P}$ that gives the length of the shortest path from a node to a dominating node, with base case: $\mathcal{L}(n, n) = 1$. When appropriate, we write $\mathcal{L}(w, n)$ to stand for the arithmetic mean over $\mathcal{L}(n', n)$ for all senses $n'$ of $w$ that are dominated by $n$.[4] In the definition of $\mathcal{S}$, $(\mathcal{N}^+)$ stands for an ordered list of nodes.

```
function SCORE ($n \in \mathcal{N}, \alpha \in \mathbb{R}^+, C \subseteq W \subseteq \mathcal{W}$) :
  $C' \leftarrow \mathcal{D}(n) \setminus C$
  return $\frac{\sum_{w \in C'} \mathcal{L}(w,n)}{|C'|^\alpha}$

function DERIVETYPES ($W \subseteq \mathcal{W}, m \in \mathbb{N}^+, p \in (0, 1]$) :
  $\alpha \leftarrow 1, C \leftarrow \{\}, R \leftarrow \{\}$
  ▷ while too few words covered
  while $|C| < p \times |W|$ :
    $n' \leftarrow \operatorname{argmin}_{n \in \mathcal{N} \setminus R} \text{SCORE}(n, \alpha, C)$
    $R \leftarrow R \cup \{n'\}$
    $C \leftarrow C \cup \mathcal{D}(n')$
    if $|R| > m$ :
      ▷ cardinality bound exceeded – restart
      $\alpha \leftarrow \alpha + \delta, C \leftarrow \{\}, R \leftarrow \{\}$
  return $R$
```

Figure 1: Algorithm for deriving slot type restrictions, with $\delta$ representing a fixed step size.

We refer to a given predicate argument position for a specified propositional template simply as a *slot*. $W \subseteq \mathcal{W}$ will stand for the set of words found to occupy a given slot (in the corpus employed), and $\mathcal{D} : \mathcal{N} \rightarrow W^*$ is a function mapping a node to the words it (partially) *sense dominates*. That is, for all $n \in \mathcal{N}$ and $w \in W$, if $w \in \mathcal{D}(n)$ then there is at least one sense $n' \in \mathcal{S}(w)$ such that $n$ is an ancestor of $n'$ as determined through use of $\mathcal{P}$. For example, we would expect the word *bank* to be dominated by a node standing for a class such as *company* as well as a separate node standing for, e.g., *location*.

Based on this model we give a greedy search algorithm in Figure 1 for deriving slot type restrictions. The algorithm attempts to find a set of dominating word senses that cover at least one of each of a majority of the words in the given set of observations. The idea is to keep the number of nodes in the dominating set small, while maintaining high coverage and not abstracting too far upward.

For a given slot we start with a set of observed words $W$, an upper bound $m$ on the number of types allowed in the result $R$, and a parameter $p$ setting a lower bound on the fraction of items in $W$ that a valid solution must dominate. For example, when $m = 3$ and $p = 0.9$, this says we require the solution to consist of no more than 3 nodes, which together must dominate at least 90% of $W$.

The search begins with initializing the cover set $C$, and the result set $R$ as empty, with the variable

[3] It is these most general templates that best correlate with existing work in verb argument preference selection; however, a given KNEXT logical form may arise from multiple distinct syntactic constructs.

[4] E.g., both senses of *female* in WN are dominated by the node for *(organism, being)*, but have different path lengths.

$\alpha$ set to 1. Observe that at any point in the execution of DeriveTypes, $C$ represents the set of all words from $W$ with at least one sense having as an ancestor a node in $R$. While $C$ continues to be smaller than the percentage required for a solution, nodes are added to $R$ based on whichever element of $\mathcal{N}$ has the smallest score.

The Score function first computes the modified coverage of $n$, setting $C'$ to be all words in $W$ that are dominated by $n$ that haven't yet been "spoken for" by a previously selected (and thus lower scoring) node. Score returns the sum of the path lengths between the elements of the modified set of dominated nodes and $n$, divided by that set's size, scaled by the exponent $\alpha$. Note when $\alpha = 1$, Score simply returns the average path length of the words dominated by $n$.

If the size of the result grows beyond the specified threshold, $R$ and $C$ are reset, $\alpha$ is incremented by some step size $\delta$, and the search starts again. As $\alpha$ grows, the function increasingly favors the coverage of a node over the summed path length. Each iteration of DeriveTypes thus represents a further relaxation of the desire to have the returned nodes be as specific as possible. Eventually, $\alpha$ will be such that the minimum scoring nodes will be found high enough in the tree to cover enough of the observations to satisfy the threshold $p$, at which point $R$ is returned.

### 4.1 Non-reliance on Frequency

As can be observed, our approach makes no use of the relative or absolute frequencies of the words in $W$, even though such frequencies could be added as, e.g., relative weights on length in Score. This is a purposeful decision motivated both by practical and theoretical concerns.

Practically, a large portion of the knowledge observed in Knext output is infrequently expressed, and yet many tend to be reasonable claims about the world (despite their textual rarity). For example, a template shown in Section 5, A __ MAY WEAR A CRASH_HELMET, was supported by just two sentences in the BNC. However, based on those two observations we were able to conclude that usually *If something wears a crash helmet, it is probably a male person.*

Initially our project began as an application of the closely related MDL approach of Li and Abe (1998), but was hindered by sparse data. We observed that our absolute frequencies were often too low to perform meaningful comparisons of relative frequency, and that different examples in development tended to call for different trade-offs between model cost and coverage. This was due as much to the sometimes idiosyncratic structure of WordNet as it was to lack of evidence.[5]

Theoretically, our goal is distinct from related efforts in acquiring, e.g., verb argument selectional preferences. That work is based on the desire to reproduce distributional statistics underlying the text, and thus relative differences in frequency are the essential characteristic. In this work we aim for general statements about the real world, which in order to gather we rely on text as a limited proxy view. E.g., given 40 hypothetical sentences supporting A MAN MAY EAT A TACO, and just 2 sentences supporting A WOMAN MAY EAT A TACO, we would like to conclude simply that A PERSON MAY EAT A TACO, remaining agnostic as to relative frequency, as we've no reason to view corpus-derived counts as (strongly) tied to the likelihood of corresponding situations in the world; they simply tell us what is generally possible and worth mentioning.

## 5 Experiments

### 5.1 Tuning to WordNet

Our method as described thus far is not tied to a particular word sense taxonomy. Experiments reported here relied on the following model adjustments in order to make use of WordNet (version 3.0).

The function $\mathcal{P}$ was set to return the union of a synset's hypernym and instance hypernym relations.

Regarding the function $\mathcal{L}$, WordNet is constructed such that always picking the first sense of a given nominal tends to be correct more often than not (see discussion by McCarthy et al. (2004)). To exploit this structural bias, we employed a modified version of $\mathcal{L}$ that results in a preference for nodes corresponding to the first sense of words to be covered, especially when the number of distinct observations were low (such as earlier, with *crash helmet*):

$$\mathcal{L}(n,n) = \begin{cases} 1 - \frac{1}{|W|} & \exists w \in W : \mathcal{S}(w) = (n, ...) \\ 1 & \text{otherwise} \end{cases}$$

[5] For the given example, this method (along with the constraints of Table 1) led to the overly general type, *living thing*.

| word | # | gloss |
|---|---|---|
| *abstraction* | 6 | a general concept formed by extracting common features from specific examples |
| *attribute* | 2 | an abstraction belonging to or characteristic of an entity |
| *matter* | 3 | that which has mass and occupies space |
| *physical entity* | 1 | an entity that has physical existence |
| *whole* | 2 | an assemblage of parts that is regarded as a single entity |

Table 1: ⟨word, sense #⟩ pairs in WordNet 3.0 considered overly general for our purposes.

| Propositional Template | Num. |
|---|---|
| A ___ CAN BE WHISKERED | 4 |
| GOVERNORS MAY HAVE ___ -S | 4 |
| A ___ CAN BE PREGNANT | 28 |
| A PERSON MAY BUY A ___ | 105 |
| A ___ MAY BARK | 6 |
| A COMPANY MAY HAVE A ___ | 713 |
| A ___ MAY SMOKE | 8 |
| A ___ CAN BE TASTY | 33 |
| A SONG MAY HAVE A ___ | 31 |
| A ___ CAN BE SUCCESSFUL | 664 |
| ___ CAN BE AT A ROAD | 20 |
| A ___ CAN BE MAGICAL | 96 |
| ___ CAN BE FOR A DICTATOR | 5 |
| ___ MAY FLOAT | 5 |
| GUIDELINES CAN BE FOR ___ -S | 4 |
| A ___ MAY WEAR A CRASH_HELMET | 2 |
| A ___ MAY CRASH | 12 |

Table 2: Development templates, paired with the number of distinct words observed to appear in the given slot.

Note that when $|W| = 1$, then $\mathcal{L}$ returns 0 for the term's first sense, resulting in a score of 0 for that synset. This will be the unique minimum, leading DERIVETYPES to act as the first-sense heuristic when used with single observations.

Parameters were set for our data based on manual experimentation using the templates seen in Table 2. We found acceptable results when using a threshold of $p = 70\%$, and a step size of $\delta = 0.1$. The cardinality bound $m$ was set to 4 when $|W| > 4$, and otherwise $m = 2$.

In addition, we found it desirable to add a few hard restrictions on the maximum level of generality. Nodes corresponding to the word sense pairs given in Table 1 were not allowed as abstraction candidates, nor their ancestors, implemented by giving infinite length to any path that crossed one of these synsets.

### 5.2 Observations during Development

Our method assumes that if multiple words occurring in the same slot can be subsumed under the same abstract class, then this information should be used to bias sense interpretation of these observed words, even when it means not picking the first sense. In general this bias is crucial to our approach, and tends to select correct senses of the words in an argument set $W$. But an example where this strategy errs was observed for the template A __ MAY BARK, which yielded the generalization that *If something barks, then it is probably a person.* This was because there were numerous textual occurrences of various types of people "barking" (speaking loudly and aggressively), and so the occurrences of *dogs* barking, which showed no type variability, were interpreted as involving the unusual sense of *dog* as a slur applied to certain people.

The template, A __ CAN BE WHISKERED, had observations including both *face* and *head*. This prompted experiments in allowing *part holonym* relations (e.g., a face is part of a head) as part of the definition of $\mathcal{P}$, with the final decision being that such relations lead to less intuitive generalizations rather than more, and thus these relation types were not included. The remaining relation types within WordNet were individually examined via inspection of randomly selected examples from the hierarchy. As with holonyms we decided that using any of these additional relation types would degrade performance.

A shortcoming was noted in WordNet, regarding its ability to represent binary valued attributes, based on the template, A __ CAN BE PREGNANT. While we were able to successfully generalize to *female person*, there were a number of words observed which unexpectedly fell outside that associated synset. For example, a *queen* and a *duchess* may each be a *female aristocrat*, a *mum* may be a *female parent*,[6] and a *fiancee* has the exclusive interpretation as being synonymous with the gender entailing *bride-to-be*.

## 6 Experiments

From the entire set of BNC-derived KNEXT propositional templates, evaluations were performed on a set of 21 manually selected examples,

[6] Serving as a good example of distributional preferencing, the primary sense of *mum* is as a flower.

| Propositional Template | Num. |
|---|---|
| A ___ MAY HAVE A BROTHER | 28 |
| A ? MAY ATTACK A ___ | 23 |
| A FISH MAY HAVE A ___ | 38 |
| A ___ CAN BE FAMOUS | 665 |
| A ? MAY ENTERTAIN A ___ | 8 |
| A ___ MAY HAVE A CURRENCY | 18 |
| A MALE MAY BUILD A ___ | 42 |
| A ___ CAN BE FAST-GROWING | 15 |
| A PERSON MAY WRITE A ___ | 47 |
| A ? MAY WRITE A ___ | 99 |
| A PERSON MAY TRY TO GET A ___ | 11 |
| A ? MAY TRY TO GET A ___ | 17 |
| A ___ MAY FALL_DOWN | 5 |
| A PERSON CAN BE HAPPY WITH A ___ | 36 |
| A ? MAY OBSERVE A ___ | 38 |
| A MESSAGE MAY UNDERGO A ___ | 14 |
| A ? MAY WASH A ___ | 5 |
| A PERSON MAY PAINT A ___ | 8 |
| A ___ MAY FLY TO A ? | 9 |
| A ? MAY FLY TO A ___ | 4 |
| A ___ CAN BE NERVOUS | 131 |

Table 3: Templates chosen for evaluation.

together representing the sorts of knowledge for which we are most interested in deriving strengthened argument type restrictions. All modification of the system ceased prior to the selection of these templates, and the authors had no knowledge of the underlying words observed for any particular slot. Further, some of the templates were purposefully chosen as potentially problematic, such as, A ? MAY OBSERVE A __, or A PERSON MAY PAINT A __. Without additional context, templates such as these were expected to allow for exceptionally broad sorts of arguments.

For these 21 templates, 65 types were derived, giving an average of 3.1 types per slot, and allowing for statements such as seen in Table 4.

One way in which to measure the quality of an argument abstraction is to go back to the underlying observed words, and evaluate the resultant sense(s) implied by the chosen abstraction. We say senses plural, as the majority of KNEXT propositions select senses that are more coarse-grained than WordNet synsets. Thus, we wish to evaluate these more coarse-grained sense disambiguation results entailed by our type abstractions.[7] We performed this evaluation using as comparisons the first-sense, and all-senses heuristics.

The first-sense heuristic can be thought of as striving for maximal specificity at the risk of precluding some admissible senses (reduced recall), while the all-senses heuristic insists on including all admissible senses (perfect recall) at the risk of including inadmissible ones.

Table 5 gives the results of two judges evaluating 314 ⟨word, sense⟩ pairs across the 21 selected templates. These sense pairs correspond to picking one word at random for each abstracted type selected for each template slot. Judges were presented with a sampled word, the originating template, and the glosses for each possible word sense (see Figure 2). Judges did not know ahead of time the subset of senses selected by the system (as entailed by the derived type abstraction). Taking the judges' annotations as the gold standard, we report precision, recall and F-score with a $\beta$ of 0.5 (favoring precision over recall, owing to our preference for *reliable* knowledge over *more*).

In all cases our method gives precision results comparable or superior to the first-sense heuristic, while at all times giving higher recall. In particular, for the case of Primary type, corresponding to the derived type that accounted for the largest number of observations for the given argument slot, our method shows strong performance across the board, suggesting that our derived abstractions are general enough to pick up multiple acceptable senses for observed words, but not so general as to allow unrelated senses.

We designed an additional test of our method's performance, aimed at determining whether the distinction between admissible senses and inadmissible ones entailed by our type abstractions were in accord with human judgement. To this end, we automatically chose for each template the observed word that had the greatest number of senses not dominated by a derived type

**A ___ MAY HAVE A BROTHER**

1 WOMAN : an adult female person (as opposed to a man); "the woman kept house while the man hunted"

2 WOMAN : a female person who plays a significant role (wife or mistress or girlfriend) in the life of a particular man; "he was faithful to his woman"

3 WOMAN : a human female employed to do housework; "the char will clean the carpet"; "I have a woman who comes in four hours a day while I write"

***4 WOMAN** : women as a class; "it's an insult to American womanhood"; "woman is the glory of creation"; "the fair sex gathered on the veranda"

Figure 2: Example of a context and senses provided for evaluation, with the fourth sense being judged as inappropriate.

[7] Allowing for multiple fine-grained senses to be judged as appropriate in a given context goes back at least to Sussna (1993); discussed more recently by, e.g., Navigli (2006).

*If something is famous, it is probably a person$_1$, an artifact$_1$, or a communication$_2$*
*If ? writes something, it is probably a communication$_2$*
*If a person is happy with something, it is probably a communication$_2$, a work$_1$, a final_result$_1$, or a state_of_affairs$_1$*
*If a fish has something, it is probably a cognition$_1$, a torso$_1$, an interior$_2$, or a state$_2$*
*If something is fast growing, it is probably a group$_1$ or a business$_3$*
*If a message undergoes something, it is probably a message$_2$, a transmission$_2$, a happening$_1$, or a creation$_1$*
*If a male builds something, it is probably a structure$_1$, a business$_3$, or a group$_1$*

Table 4: Examples, both good and bad, of resultant statements able to be made post-derivation. Authors manually selected one word from each derived synset, with subscripts referring to sense number. Types are given in order of support, and thus the first are examples of "Primary" in Table 5.

| Method | $\bigcup_j$ | | | $\bigcap_j$ | | | Type |
|---|---|---|---|---|---|---|---|
| | Prec | Recall | F$_{.5}$ | Prec | Recall | F$_{.5}$ | |
| derived | 80.2 | 39.2 | **66.4** | 61.5 | 47.5 | **58.1** | All |
| first | 81.5 | 28.5 | 59.4 | 63.1 | 34.7 | 54.2 | |
| all | 59.2 | 100.0 | 64.5 | 37.6 | 100.0 | 42.9 | |
| derived | 90.0 | 50.0 | **77.6** | 73.3 | 71.0 | **72.8** | Primary |
| first | 85.7 | 33.3 | 65.2 | 66.7 | 45.2 | 60.9 | |
| all | 69.2 | 100.0 | 73.8 | 39.7 | 100.0 | 45.2 | |

Table 5: Precision, Recall and F-score ($\beta = 0.5$) for coarse grained WSD labels using the methods: derive from corpus data, first-sense heuristic and all-sense heuristic. Results are calculated against both the union $\bigcup_j$ and intersection $\bigcap_j$ of manual judgements, calculated for all derived argument types, as well as Primary derived types exclusively.

THE STATEMENT ABOVE IS A REASONABLY CLEAR, ENTIRELY PLAUSIBLE GENERAL CLAIM AND SEEMS NEITHER TOO SPECIFIC NOR TOO GENERAL OR VAGUE TO BE USEFUL:

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

Figure 3: Instructions for evaluating KNEXT propositions.

restriction. For each of these alternative (non-dominated) senses, we selected the ancestor lying at the same distance towards the root from the given sense as the average distance from the dominated senses to the derived type restriction. In the case where going this far from an alternative sense towards the root would reach a path passing through the derived type and one of its subsumed senses, the distance was cut back until this was no longer the case.

These alternative senses, guaranteed to not be dominated by derived type restrictions, were then presented along with the derived type and the original template to two judges, who were given the same instructions as used by Van Durme and Schubert (2008), which can be found in Figure 3.

| | judge 1 | judge 2 | corr |
|---|---|---|---|
| derived | 1.76 | 2.10 | 0.60 |
| alternative | 3.63 | 3.54 | 0.58 |

Table 6: Average assessed quality for derived and alternative synsets, paired with Pearson correlation values.

Results for this evaluation are found in Table 6, where we see that the automatically derived type restrictions are strongly favored over alternative abstracted types that were possible based on the given word. Achieving even stronger rejection of alternative types would be difficult, since KNEXT templates often provide insufficient context for full disambiguation of all their constituents, and judges were allowed to base their assessments on any interpretation of the verbalization that they could reasonably come up with.

## 7 Related Work

There is a wealth of existing research focused on learning probabilistic models for selectional restrictions on syntactic arguments. Resnik (1993) used a measure he referred to as *selectional preference strength*, based on the KL-divergence between the probability of a class and that class given a predicate, with variants explored by Ribas (1995). Li and Abe (1998) used a *tree cut* model over WordNet, based on the principle of *Minimum Description Length* (MDL). McCarthy has performed extensive work in the areas of selectional preference and WSD, e.g., (McCarthy, 1997; McCarthy, 2001). Calling the generalization problem a case of engineering in the face of sparse data, Clark and Weir (2002) looked at a number of previous methods, one conclusion being that the approach of Li and Abe appears to over-generalize.

Cao et al. (2008) gave a distributional method for deriving semantic restrictions for FrameNet frames, with the aim of building an Italian FrameNet. While our goals are related, their work can be summarized as taking a pre-existing gold standard, and extending it via distributional similarity measures based on shallow contexts (in this case, $n$-gram contexts up to length 5). We have presented results on strengthening type restrictions on arbitrary predicate argument structures derived directly from text.

In describing ALICE, a system for *lifelong learning*, Banko and Etzioni (2007) gave a summary of a proposition abstraction algorithm developed independently that is in some ways similar to DERIVETYPES. Beyond differences in node scoring and their use of the first sense heuristic, the approach taken here differs in that it makes no use of relative term frequency, nor contextual information outside a particular propositional template.[8] Further, while we are concerned with general knowledge acquired over diverse texts, ALICE was built as an agent meant for constructing domain-specific theories, evaluated on a 2.5-million-page collection of Web documents pertaining specifically to nutrition.

Minimizing word sense ambiguity by focusing on a specific domain was later seen in the work of Liakata and Pulman (2008), who performed hierarchical clustering using output from their KNEXT-like system first described in (Liakata and Pulman, 2002). Terminal nodes of the resultant structure were used as the basis for inferring semantic type restrictions, reminiscent of the use of CBC clusters (Pantel and Lin, 2002) by Pantel et al. (2007), for typing the arguments of paraphrase rules.

Assigning pre-compiled instances to their first-sense reading in WordNet, Paşca (2008) then generalized *class attributes* extracted for these terms, using as a resource Google search engine query logs.

Katrenko and Adriaans (2008) explored a constrained version of the task considered here. Using manually annotated semantic relation data from SemEval-2007, pre-tagged with correct argument senses, the authors chose the least common subsumer for each argument of each relation considered. Our approach keeps with the intuition of preferring specific over general concepts in WordNet, but allows for the handling of relations automatically discovered, whose arguments are not pre-tagged for sense and tend to be more wide-ranging. We note that the least common subsumer for many of our predicate arguments would in most cases be far too abstract.

## 8 Conclusion

As the volume of automatically acquired knowledge grows, it becomes more feasible to abstract from existential statements to stronger, more general claims on what usually obtains in the real world. Using a method motivated by that used in deriving selectional preferences for verb arguments, we've shown progress in deriving semantic type restrictions for arbitrary predicate argument positions, with no prior knowledge of sense information, and with no training data other than a handful of examples used to tune a few simple parameters.

In this work we have made no use of relative term counts, nor corpus-wide, distributional frequencies. Despite foregoing these often-used statistics, our methods outperform abstraction based on a strict first-sense heuristic, employed in many related studies.

Future work may include a return to the MDL approach of Li and Abe (1998), but using a frequency model that "corrects" for the biases in texts relative to world knowledge – for example, correcting for the preponderance of people as subjects of textual assertions, even for verbs like *bark, glow*, or *fall*, which we know to be applicable to numerous non-human entities.

**Acknowledgements** Our thanks to Matthew Post and Mary Swift for their assistance in evaluation, and Daniel Gildea for regular advice. This research was supported in part by NSF grants IIS-0328849 and IIS-0535105, as well as a University of Rochester Provost's Multidisciplinary Award (2008).

[8] Banko and Etzioni abstracted over subsets of pre-clustered terms, built using corpus-wide distributional frequencies

## References

Michele Banko and Oren Etzioni. 2007. Strategies for Lifelong Knowledge Extraction from the Web. In *Proceedings of K-CAP*.

BNC Consortium. 2001. The British National Corpus, version 2 (BNC World). Distributed by Oxford University Computing Services.

Diego De Cao, Danilo Croce, Marco Pennacchiotti, and Roberto Basili. 2008. Combining Word Sense and Usage for Modeling Frame Semantics. In *Proceedings of Semantics in Text Processing (STEP)*.

Eugene Charniak. 2000. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL*.

Stephen Clark and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2).

Michael Collins. 1997. Three Generative, Lexicalised Models for Statistical Parsing. In *Proceedings of ACL*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Sophia Katrenko and Pieter Adriaans. 2008. Semantic Types of Some Generic Relation Arguments: Detection and Evaluation. In *Proceedings of ACL-HLT*.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2).

Maria Liakata and Stephen Pulman. 2002. From Trees to Predicate Argument Structures. In *Proceedings of COLING*.

Maria Liakata and Stephen Pulman. 2008. Automatic Fine-Grained Semantic Classification for Domain Adaption. In *Proceedings of Semantics in Text Processing (STEP)*.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Using automatically acquired predominant senses for Word Sense Disambiguation. In *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.

Diana McCarthy. 1997. Estimation of a probability distribution over a hierarchical classification. In *The Tenth White House Papers COGS - CSRP 440*.

Diana McCarthy. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. Ph.D. thesis, University of Sussex.

Roberto Navigli. 2006. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of COLING-ACL*.

Marius Paşca. 2008. Turning Web Text and Search Queries into Factual Knowledge: Hierarchical Class Attribute Extraction. In *Proceedings of AAAI*.

Patrick Pantel and Dekang Lin. 2002. Discovering Word Senses from Text. In *Proceedings of KDD*.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning Inferential Selectional Preferences. In *Proceedings of NAACL-HLT*.

Philip Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*

Francesc Ribas. 1995. On learning more appropriate Selectional Restrictions. In *Proceedings of EACL*.

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

Lenhart K. Schubert and Matthew H. Tong. 2003. Extracting and evaluating general world knowledge from the brown corpus. In *Proceedings of HLT/NAACL Workshop on Text Meaning*, May 31.

Lenhart K. Schubert. 2002. Can we derive general world knowledge from texts? In *Proceedings of HLT*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In *Proceedings of WWW*.

Michael Sussna. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of CIKM*.

Benjamin Van Durme and Lenhart Schubert. 2008. Open Knowledge Extraction through Compositional Language Processing. In *Proceedings of Semantics in Text Processing (STEP)*.

Benjamin Van Durme, Ting Qian, and Lenhart Schubert. 2008. Class-driven Attribute Extraction. In *Proceedings of COLING*.