Class-Driven Attribute Extraction

Benjamin Van Durme, **Ting Qian** and **Lenhart Schubert**
Department of Computer Science
University of Rochester
Rochester, NY 14627, USA

Abstract

We report on the large-scale acquisition of class attributes with and without the use of lists of representative instances, as well as the discovery of *unary* attributes, such as typically expressed in English through prenominal adjectival modification. Our method employs a system based on compositional language processing, as applied to the British National Corpus. Experimental results suggest that document-based, open class attribute extraction can produce results of comparable quality as those obtained using web query logs, indicating the utility of exploiting explicit occurrences of class labels in text.

1 Introduction

Recent work on the task of acquiring attributes for concept classes has focused on the use of pre-compiled lists of class representative instances, where attributes recognized as applying to multiple instances of the same class are inferred as being likely to apply to most, or all, members of that class. For example, the class *US President* might be represented as a list containing the entries *Bill Clinton*, *George Bush*, *Jimmy Carter*, etc. Phrases such as *Bill Clinton's chief of staff ...*, or search queries such as *chief of staff bush*, provide evidence that the class *US President* has as an attribute *chief of staff*.

Usually the focus of such systems has been on on *binary* attributes, such as the example *chief of staff*, while less attention has been paid to *unary* class attributes such as *illegal* for the class *Drug*, or *warm-blooded* for the class *Animal*.[1] These attributes are most typically expressed in English through prenominal adjectival modification, with the nominal serving as a class designator. When attribute extraction is based entirely on instances and not the class labels themselves, this form of modification goes undiscovered.

In what follows we explore both the impact of gazetteers in attribute extraction as well as the acquisition and filtering of unary class attributes, through a process based on logical form generation from syntactic parses derived from the British National Corpus.

2 Extraction Framework

Extraction was performed using a modified version of the KNEXT system, a knowledge acquisition framework constructed for large scale generation of abstracted logical forms through compositional linguistic analysis. The following provides an overview of KNEXT and its target knowledge representation, Episodic Logic.

2.1 Episodic Logic

Automatically acquiring general world knowledge from text is not a task that provides an immediate solution to any real world problem.[2] Rather, the motivation for acquiring large stores of background knowledge is to enable research within other areas of artificial intelligence, e.g., the construction of systems that can engage in dialogues about everyday topics in unrestricted English, use

[1] Almuhareb and Poesio (2004) treat unary attributes as values of binary attributes; e.g., *illegal* might be the value of a *legality* attribute. But for many unary attributes, this is a stretch.

[2] Unless one regularly needs reminding of facts such as, A WOMAN MAY BOIL A GOAT.

```
(Some e0:
    [e0 at-about Now0]
    [(Many.det x :
        [x ((attr athletic.a) (plur youngster.n))]
        [x want.v
            (Ka
              (become.v
                 (plur
                    ((attr professional.a) athlete.n)))))])
     ** e0])
```

Figure 1: Example EL formula; square brackets indicate a sentential infix syntax of form [*subject pred object ...*], *Ka* reifies action predicates, and *attr* "raises" adjectival predicates to predicate modifiers; *e0* is the situation characterized by the sentence.

common sense in answering questions or solving problems, pursue intrinsic goals independently, and show awareness of their own characteristics, biography, and cognitive capacities and limitations. An important challenge in the pursuit of these long-range goals is the design and implementation of a knowledge representation that is as expressively rich as natural language and facilitates language understanding and commonsense reasoning.

Episodic Logic (EL) (Schubert and Hwang, 2000), is a superset of FOL augmented with certain semantic features common to all human languages: generalized quantification, intensionality, uncertainty, modification and reification of predicates and propositions, and event characterization by sentences. An implementation of EL exists as the EPILOG system (Schaeffer et al., 1993), which supports both forward and backward inference, along with various specialized routines for dealing with, e.g., color, time, class subsumption, etc. EPILOG is under current development as a platform for studying a notion of explicit self-awareness as defined by Schubert (2005).

As an indication of EL's NL-like syntax, figure 1 contains the output of EPILOG's parser/logical-form generator for the sentence, *Many athletic youngsters want to become professional athletes*.

2.2 KNEXT

If "deep" language understanding and commonsense reasoning involve items as complex and structured as seen in figure 1, then automated knowledge acquisition cannot simply be a matter of accumulating rough associations between word strings, along the lines "(Youngster) (want become) (professional athlete)". Rather, acquired knowledge needs to conform with a systematic, highly expressive KR syntax such as EL.

The KNEXT project is aimed at extracting such structured knowledge from text. One of the major obstacles is that the bulk of commonsense knowledge on which people rely is not explicitly written down – precisely because it is *common*. Even it were written down, most of it could not be reliably interpreted, because reliable interpretation of language is itself dependent on commonsense knowledge (among other things).

In view of these difficulties, KNEXT has initially focused on attempting to abstract world knowledge "factoids" from texts, based on the logical forms derived from parsed sentences. The idea is that nominal pre- and post-modifiers, along with subject-verb-object relations, captured in logical forms similar to that in figure 1, give a glimpse of the common properties and relationships in the world – even if the source sentences describe invented situations. For example, the following were extracted by KNEXT, then automatically verbalized back into English for ease of readability:

- SOME_NUMBER_OF YOUNGSTERS MAY WANT TO BECOME ATHLETES.
- YOUNGSTERS CAN BE ATHLETIC.
- ATHLETES CAN BE PROFESSIONAL.

2.3 Attribute Extraction via KNEXT

In order to study the contribution of lists of instances (i.e., generalized *gazetteers*) to the task of attribute extraction, the version of KNEXT as presented by Schubert (2002) was modified to provide output of a form similar to that of the extraction work of Paşca and Van Durme (2007).

KNEXT's abstracted, propositional output was automatically verbalized into English, with any resultant statements of the form, *A(N) X MAY HAVE A(N) Y*, taken to suggest that the class *X* has as an attribute the property *Y*.

KNEXT was designed from the beginning to make use of gazetteers if available, where a phrase such as *Bill Clinton vetoed the bill* supports the (verbalized) proposition *A PRESIDENT MAY VETO A BILL.* just as would *The president vetoed the bill.* We instrumented the system to record which propositions did or did not require gazetteers in their construction, allowing for a numerical breakdown of the respective contributions of known instances of a class, versus the class label itself.

Paşca and Van Durme (2007) described the results of an informal survey asking participants to enumerate what they felt to be important attributes for a small set of example classes. Some of these resultant attributes were not of the form targeted by the authors' system. For example, *nonprofit* was given as an important potential attribute for the class *Company*, as well as *legal* for the class *Drug*. These attributes correspond to *unary* predicates as compared to the targeted *binary* predicates underlying such attributes as *cost(X,Y)* for the class *Drug*.

We extracted such unary attributes by focusing on verbalizations of the form, *A(N) X CAN BE Y* as in *AN ANIMAL CAN BE WARM-BLOODED*.

3 Experimental Setting

3.1 Corpus Processing

Initial reports on the use of KNEXT were focused on the processing of manually created parse trees, on a corpus of limited size (the Brown corpus of Kucera and Francis (1967)). Since that time the system has been modified into a fully automatic extraction system, making use of syntactic parse trees generated by parsers trained on the Penn Treebank.

For our studies here, the parser employed was that of Collins (1997) applied to the sentences of the British National Corpus (BNC Consortium, 2001). Our choice of the BNC was motivated by its breadth of genre, its substantial size (100 million words) and its familiarity (and accessibility) to the community.

3.2 Gazetteers

KNEXT's gazetteers were used as-is, and which were defined based on a variety of sources; miscellaneous publicly available lists, as well as manual enumeration. The classes covered can be seen in the Results section in table 2, where the minimum, maximum and mean size were 2, 249, and 41, respectively.

3.3 Filtering out Non-predicative Adjectives

Beyond the pre-existing KNEXT framework, additional processing was introduced for the extraction of unary attributes in order to filter out vacuous or unsupported propositions derived from non-compositional phrases.

This filtering was performed through the creation of three lists: a whitelist of accepted predicative adjectives; a graylist containing such adjectives that are meaningful as unary predicates only when applied to plural nouns; and a blacklist derived from Wikipedia topic titles, representing lexicalized, non-compositional phrases.

Whitelist The creation of the whitelist began with calculating part-of-speech (POS) tagged bigram counts using the Brown corpus. The advantage of using a POS-tagged bigram model lies in the saliency of phrase structures, which enabled frequency calculations for both attributive and predicative uses of a given adjective. Attributive counts were based on instances when an adjective appears in the pre-nominal position and modifies another noun. Predicative counts were derived by summing over occurrences of a given adjective after all possible copulas. These counts were used to compute a p/a ratio - the quotient of predicative count over attributive count - for each word classified by WordNet (Fellbaum, 1998) as a having an adjectival use. After manual inspection, two cut-off points were chosen at ratios of .06 and 0, as seen in table 1.

Words not appearing in the Brown corpus (i.e. having 0 count for both uses), were sampled and inspected, with the decision made to place the majority within the whitelist, excluding just those with suffixes including -al, -c, -an, -st, -ion, -th, -o, -ese, -er, -on, -i, -x, -v, and -ing.

This process resulted in a combined whitelist of 14,249 (usually) predicative adjectives.

p/a ratio (r)	**Cut-off decision**
$r \geq .06$	keep the adjective*
$0 < r < .06$	remove the adjective*
otherwise	keep the adjective*

Table 1: Cut-off decision given the p/a ratio of an adjective. **Note: except for hand-selected cases.*

Graylist We manually constructed a short list (currently 33 words) containing adjectives that are generally inappropriate as whitelist entries, but could be acceptable when applied to plurals. For example, the verbalized proposition *OBJECTS CAN BE SIMILAR* was deemed acceptable, while a statement such as *AN OBJECT CAN BE SIMILAR* is erroneous due to a missing argument.

Blacklist From an exhaustive set of Wikipedia topic titles was derived a blacklist consisting of entries that had to satisfy four criteria: 1) no more than three words in length; 2) has no closed-class

words, such as prepositions or adverbs; 3) must begin with an adjective and end with a noun (determined by WordNet); and 4) does not contain any numerical characters or miscellaneous symbols that are usually not meaningful in English. Therefore, each title in the resultant list is literally a short noun phrase with adjectives as premodifiers. It was observed that in these encyclopedia titles, the role of adjectives is predominantly to restrict the scope of the object that is being named (e.g. CRIMINAL LAW), rather than to describe its attributions or features (e.g. DARK EYES). More often than not, only cases similar to the second example can be safely verbalized as *X CAN BE Y* from a noun phrase *Y X*, with Y being the pre-nominal adjective.

We further refined this list by examining trigram frequencies as reported in the web-derived n-gram collection of Brants and Franz (2006). For each title of the form *(Adj N)* ..., we gathered trigram frequencies for adverbial modifications such as *(very Adj N)* ..., and *(truly Adj N)* .... Intuitively, high relative frequency of such modification with respect to the non-modified bigram supports removal of the given title from the blacklist.

Trigram counts were collected using the modifiers: *absolutely*, *almost*, *entirely*, *highly*, *nearly*, *perfectly*, *truly* and *very*. These counts were summed for a given title then divided by the aforementioned bigram score. Upon sampled inspection, all three-word titles were kept on the blacklist, along with any two-word title with a resultant ratio less than 0.028. For example, the titles *Hardy Fish*, *Young Galaxy*, and *Sad Book* were removed, while *Common Cause*, *Bouncy Ball*, and *Heavy Oil* were retained.

4 Results

From the parsed BNC, 6,205,877 propositions were extracted, giving an average of 1.396 propositions per input sentence.[3] These results were then used to explore the necessity of gazetteers, and the potential for extracting unary attributes. Quality judgements were performed using a 5 point scale as seen in figure 2.

[3]These approximately six million verbalized propositions, along with their underlying logical form and respective source sentence(s), may be browsed interactively through an online browser available at: `http://www.cs.rochester.edu/u/vandurme/epik`

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 2: Instructions for scaled judging.

4.1 Necessity of Gazetteers

From the total set of extracted propositions, 638,809 could be verbalized as statements of the form *X MAY HAVE Y*. There were 71,531 unique classes (*X*) for which at least a single candidate attribute (*Y*) was extracted, with 9,743 of those having at least a single such attribute that was supported by a minimum of two distinct sentences.

Table 2 gives the number of attributes extracted for the given classes when using only gazetteers, when using only the given names as class labels, and when using both together. While instance-based extraction generated more unique attributes, there were still a significant number of results derived based exclusively on class labels. Further, as can be seen for cases such as *Artist*, class-driven extraction provided a large number of attribute candidates not observed when relying only on gazetteers (701 total candidate attributes were gathered based on the union of 441 and 303 candidates respectively extracted with, and without a gazetteer for *Artist*).

We note that this volume measure is potentially biased against class-driven extraction, as no effort was made to pick an optimal label for a given gazetteer, (the original hand-specified class labels were retained). For example, one might expect the label *Drink* to generate more, yet still appropriate, propositions than *Beverage*, *Actor* and/or *Actress* as compared to *Show Biz Star*, or the semantically similar *Book* versus *Literary Work*. This is suggested by the entries in the table based on using supertypes of the given class, as well as in figure 3, which favorably compares top attributes discovered for select classes against those reported elsewhere in the literature.

Table 3 gives the assessed quality for the top ten attributes extracted for five of the classes in table 2. As can be seen, class-driven extraction can produce attributes of quality assessed at par with attributes extracted using only gazetteers.

BasicFood	Religion
K (*Food*): quality, part, taste, value, portion.. D: species, pounds, cup, kinds, lbs, bowl.. Q: nutritional value, health benefits, glycemic index, varieties, nutrition facts, calories..	K: basis, influence, name, truths, symbols, principles, strength, practice, origin, adherent, god, defence.. D: teachings, practice, beliefs, religion spread, principles, emergence, doctrines.. Q: basic beliefs, teachings, holy book, practices, rise, branches, spread, sects..
HeavenlyBody	**Painter**
K_G (*Planet*): surface, orbit, bars, history, atmosphere.. K (*Planet*): surface, history, future, orbit, mass, field.. K (*Star*): surface, mass, field, regions.. D: observations, spectrum, planet, spectra, conjunction, transit, temple, surface.. Q: atmosphere, surface, gravity, diameter, mass, rotation, revolution, moons, radius..	K_G (*Artist*) : works, life, career, painting, impression, drawings, paintings, studio, exhibition.. K (*Artist*): works, impression, career, life, studio.. K (*Painter*) : works, life, wife, eye.. Q': paintings, works, portrait, death, style, artwork, bibliography, bio, autobiography, childhood..

Figure 3: Qualitative comparison of top extracted attributes; K_G is KNEXT using gazetteers, K (*class*) is KNEXT for a class label similar to the heading, D and Q are document- and query-based results as reported in (Paşca et al., 2007), Q' is query-based results reported in (Paşca and Van Durme, 2007).

The noticeable drop in quality for the class *Planet* when only using gazetteers (3.2 mean judged acceptability) highlights the recurring problem of word sense ambiguity in extraction. The names of Roman deities, such as Mars or Mercury, are used to refer to a number of conceptually distinct items, such as planets within our solar system. Two of the attributes judged as poor quality for this class were *bars* and *customers*, respectively derived from the noun phrases: (NP (NNP Mars) (NNS bars)), and (NP (NNP Mercury) (NNS customers)). Note that in both cases the underlying extraction is correctly performed; the error comes from abstracting to the wrong class. These NPs may arguably support the verbalized propositions, e.g.: A CANDY-COMPANY MAY HAVE BARS, and A CAR-COMPANY MAY HAVE CUSTOMERS.

These examples point to additional areas for improvement beyond sense disambiguation: non-compositional phrase filtering for all NPs, rather than just in the cases of adjectival modification (*Mars bar* is a Wikipedia topic); and relative discounting of patterns used in the extraction process[4]. This later technique is commonly used in specialized extraction systems, such as constructed by Snow et al. (2005) who fit a logistic regression model for hypernym (X is-a Y) classification based on WordNet, and Girju et al. (2003) who trained a classifier to look specifically for part-whole relations.

[4]For example, (NP (NNP X) (NNS Y)) may be more semantically ambiguous than, e.g., the possessive construction (NP (NP (NNP X) (POS 's)) (NP (NNS Y))).

4.2 Unary Attributes

Table 4 shows how filtering non-compositional phrases from CAN BE propositions affects extraction volume. Table 5 shows the difference between such post-filtered propositions and those that were deleted. As our filter lists were not built fully automatically, evaluation was performed exclusively by an author with negligible direct involvement in the lists' creation (so-as to minimize judgement bias).

As examples, the top ten unary attributes for select classes are given in table 6, which the authors believe to be high quality on average, with some bad entries present. Attributes such as *pre-raphaelite* for *Painter* are considered obscure, while those such as *favourite* for *Animal* are considered unlikely to be useful as a unary predicate.

The importance of class-driven extraction can be seen in results such as those given for the class *Apple*. Even if it were the case that gazetteer-based extraction could deliver perfect results for those classes whose instances occasionally appear explicitly in text, there are a number of classes for which such instances are entirely lacking. For example, there are many instances of the class *Company* which have been individually named and appear in text with some frequency, e.g., Microsoft, Walmart, or Boeing. However, despite the many real-world instantiations of the class *Apple*, this does not translate into a list of individually named members in text.[5] If our goal is to acquire attributes for as many classes as possible, our results

[5]Instances of *Apple* are referred to directly as such; "an apple."

Class	Both	Gaz.	Class Lbl.
Continent	777	698	96
Country	7,285	5,993	1,696
US State	1,289	1,286	609*
US City	2,216	2,120	813*
World City	4,780	4,747	813*
Beverage	53	53	0
Tycoon	19	10	10
TV Network	71	71	0
Artist	706	441	303
Medicine	29	2	27
Weekday	1,234	1,232	2
Month	2,282	1,875	474
Dictator	533	509	28
Conqueror	103	84	19
Philosopher	672	649	37
Conductor	118	74	45
Singer	220	179	49
Band	349	58	303
King	811	208	664
Queen	541	17	532
Religious Leader	127	127	0
Adventurer	32	27	5
Planet	289	163	141
Criminal/Outlaw	30	30	6/4*
Service Agency	85	83	2
Architect	72	67	63
Show Biz Star	82	82	0
Film Maker	42	33	9
Composer	722	651	98
Humanitarian	5	5	0
Pope	235	123	113
River	402	168	253
Company	3,968	1,553	2,941
Deity	1,037	1,027	19
Scientist	798	750	60
Religious Holiday	594	593	65*
Civic Holiday	3	3	65*
Military Commander	71	71	26*
Intl Political Entity	673	673	0
Sports Celebrity	45	45	0
Activist Organization	63	63	0
Martial Art	3	3	0
Government Agency	295	294	2
Criminal Organization	0	0	0
US President	596	596	1,421*
Political Leader	568	568	170*
Supreme Court Justice	0	0	18*
Emperor	436	211	259
Fictitious Character	227	227	180*
Literary Work	9	9	0
Engineer/Inventor	10	10	73/13*
Famous Lawyer	0	0	72*
Writer	1,116	957	236
TOTAL	35,723	29,518	8,506

Table 2: Extraction volume with and without using gazetteers. **Note: When results are zero after gaz. omission, values are reported for super-types, such as* Holiday *for the sub-type* Civic Holiday*, or* City *for* US City*. A/B scores reported for each class used separately, e.g.,* Engineer/Inventor.

Class	Both	Gazetteer	Class Label
King	1.2	1.9	1.3
Composer	1.5	1.5	2.1
River	1.9	1.9	1.5
Continent	1.5	1.9	2.0
Planet	1.9	**3.2**	1.6

Table 3: Average judged acceptability for the top ten attributes extracted for the given classes when using/not-using gazetteer information.

Collection	**Size**	**% of**	
		Original	CAN BE
Original total	6,204,184	100	-
Filtered total	5,382,282	87	-
Original CAN BE	2,895,325	46	100
Filtered CAN BE	2,073,417	33	72
Whitelist	812,146	15	28
Blacklist	19,786	1<	1

Table 4: Impact of filtering on volume. For example, those propositions removed because of the whitelist comprised 15% of the total propositions extracted, or 28% of those specifically verbalized as *X CAN BE Y*.

indicate the benefits of exploiting the explicit appearance of class labels in text.

5 Related Work

Paşca and Van Durme (2007) presented an approach to attribute extraction based on the use of search engine query logs, a previously unexplored source of data within information extraction. Results confirmed the intuition that a significant number of high quality, characteristic attributes for many classes may be derived based on the relative frequency with which anonymous users request particular pieces of information for known instances of a concept class. Paşca et al. (2007) compared the quality of shallow attribute extraction techniques as applied to documents versus search engine query logs, concluding that such methods are more applicable to query logs than to documents. We note that while search queries do seem ideally suited for extracting class attributes, existing large-scale collections of query logs are proprietary and thus unavailable to the general research community. At least until such a resource becomes available, it is of interest to the community that (qualitatively) similar extraction results may be achieved exclusively using publicly available document collections.

Alternative approaches to harvesting large-scale knowledge repositories based on logical forms include that reported by Suchanek et al. (2007). The authors used non-linguistic information avail-

	1	10	100	1,000
Filtered	3.18	3.60	2.74	2.76
Blacklist	3.88	4.00	4.08	4.06
Whitelist	3.78	3.76	3.74	3.80

Table 5: Mean evaluated acceptability for 50 unary attributes randomly sampled from each of the given levels of support (attribute occurred once, less than 10 times, less than 100 times, ...). Filtered refers to the final "clean" results, Blacklist and Whitelist refer to propositions deleted due to the given list.

Painter	famous, romantic, distinguished, celebrated, well-known, pre-raphaelite, flemish, dutch, abstract
Animal	dead, trapped, dangerous, unfortunate, intact, hungry, wounded, tropical, sick, favourite
Drug	dangerous, powerful, addictive, safe, illegal, experimental, effective, prescribed, harmful, hallucinatory
Apple	red, juicy, fresh, bad, substantive, stuffed, shiny, ripe, green, baked
Earthquake	disastrous, violent, underwater, prolonged, powerful, popular, monstrous, fatal, famous, epic

Table 6: Top ten unary attributes for select classes, gathered exclusively without the use of gazetteers.

able via Wikipedia to populate a KB based on a variant of the logic underlying the Web Ontology Language (OWL). Results were limited to 14 predefined relation types, such as *diedInYear* and *politicianOf*, with membership of instances within particular concept classes inferred based on Wikipedia's category pages. Authors report 5 million so-called *ontological* facts being extracted.

Almuhareb and Poesio (2004) performed attribute extraction on webtext using simple extraction patterns (e.g., "the * of the C [is|was]", and "[a|an|the] * C [is|was]", which respectively match *The color of the rose was red* and *A red rose was* ...), and showed that such attributes could improve concept clustering. Subsequently they tested an alternative approach to the same problem using a dependency parser, extracting syntactic relations such as (ncmod, rose, red) and (ncsubj, grow, rose) (Almuhareb and Poesio,). They concluded that syntactic information is relatively expensive to derive, and serves primarily to alleviate data sparsity problems (by capturing dependencies between potentially widely separated words) that may no longer be an issue given the scale of the Web. We take a different view, first because attribute extraction is an offline task for which a 60% overhead cost (reported by the authors) is not a major issue, but more importantly because we regard approaches that process language compositionally as ultimately necessary for deeper meaning representation and language understanding.

Following intuitions similar to those laid out by Schubert (2002), Banko et al. (2007) presented TextRunner, the latest in a series of ever more sophisticated general information extraction systems (Cafarella et al., 2005; Etzioni et al., 2004). The authors constructed a non-parser based extractor for open domain text designed to efficiently process web-sized datasets. Results are in the form of bracketed text sequences that hint at a sentence's underlying semantics. For example, *(Bletchley Park) was location of (Station X)*.

Cimiano et al. (2005) performed a limited form of class-driven extraction in order to induce class hierarchies via the methods of Formal Concept Analysis (FCA). For example, a *car* is both *driveable* and *rentable* based on its occurrence in object position of the relevant verbs. A *bike* shares these properties with *car*, as well as having the property *rideable*, leading to these classes being near in the resultant automatically constructed taxonomy. Experiments were performed on limited domains for which pre-existing ontologies existed for measuring performance (*tourism* and *finance*).

Lin (1999) gave a corpus-based method for finding various types of non-compositional phrases, including the sort discussed in this paper. Identification was based on mutual information statistics conditioned on a given syntactic context (such as our targeted prenominal adjectival modification). If the mutual information of, e.g., *white house*, shows strong differences from that for constructions with similar components, e.g., *red house*, and *white barn*, then the given phrase was determined to be non-compositional. The use of this method to supplement that explored here is a matter of current investigation. Early results confirm our intuition regarding the correlation between such automatically discovered non-compositional phrases and Wikipedia topic titles, where high scoring phrases not already in our list tend to suggest miss-

Yes	cooking pot, magic flute, runny nose, skimmed milk, acquired dyslexia, charged particles, earned income
No	causal connectives, golden oldies, ruling junta, graduated pension, unsung heroes, viral rna

Table 7: Example high-scoring phrases as ranked by Lin's metric when applied to KNEXT logical forms, along with whether there is, at the time of this writing, an associated Wikipedia entry.

A CAR MAY HAVE A ...
back, boot, side, driver, front, roof, seat, end, interior, owner, door, control, value, bonnet, wheel, window, engine, headlights..
A CAR CAN BE ...
black, parked, red, white, armoured, nice, hired, bloody, open, beautiful, wrecked, unmarked, secondhand, powerful, brand-new, out-of use, damaged, heavy, dark, competitive, broken-down..
A CAR MAY BE ... IN SOME WAY
parked, stolen, driven, damaged, serviced, stopped, lost, clamped, overturned, locked, involved in an accident, found, turned, transported..

Table 8: Top attributes extracted for the class *Car*, where MAY BE relational properties (akin to those used by Cimiano et al. (2005)) are similarly acquired via verbalization of abstracted logical forms.

ing entries in the encyclopedia (see table 7). The ability to perform such "missing topic discovery" should be of interest to those within the emerging community of Wikipedia-focused AI researchers.

6 Conclusion

We have shown that an open knowledge extraction system can effectively yield class attributes, even when named instances of the class are unavailable or scarce (as a final example see table 8). We studied the quantitative contributions of instances (as given in KNEXT gazetteers) and explicitly occurring class nominals to the discovery of attributes, and found both to be important. We paid particular attention to the acquisition of *unary* class attributes, for which access to class labels is of particular importance because of their typical manner of expression in text.

Acknowledgements The authors are grateful to Daniel Gildea for contributing a parsed version of the BNC. This work was supported by NSF grants IIS-0328849 and IIS-0535105.

References

Almuhareb, Abdulrahman and Massimo Poesio. Finding concept attributes in the web using a parser. In *Proceedings Corpus Linguistics Conference*.

Almuhareb, Abdulrahman and Massimo Poesio. 2004. Attribute-based and value-based clustering: an evaluation. In *Proceedings of EMNLP*.

Banko, Michele, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web. In *Proceedings of IJCAI*.

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

Brants, Thorsten and Alex Franz. 2006. Web 1T 5-gram Version 1. Distributed by the Linguistic Data Consortium.

Cafarella, Michael J., Doug Downey, Stephen Soderland, and Oren Etzioni. 2005. KnowItNow: Fast, Scalable Information Extraction from the Web. In *Proceedings of HLT-EMNLP*.

Cimiano, Philipp, Andreas Hotho, and Steffen Stabb. 2005. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Inteligence Research*.

Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of ACL*.

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 *Proceedings of WWW*.

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

Girju, R., A. Badulescu, and D. Moldovan. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of HLT-NAACL*.

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

Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of ACL*.

Paşca, Marius and Benjamin Van Durme. 2007. What you seek is what you get: Extraction of class attributes from query logs. In *Proceedings of IJCAI*.

Paşca, Marius, Benjamin Van Durme, and Nikesh Garera. 2007. The role of documents vs. queries in extracting class attributes from text. In *Proceedings of CIKM*.

Schaeffer, S.A., C.H. Hwang, J. de Haan, and L.K. Schubert. 1993. EPILOG, the computational system for episodic logic: User's guide. Technical report, Dept. of Computing Science, Univ. of Alberta, August.

Schubert, Lenhart K. and Chung Hee Hwang. 2000. Episodic logic meets little red riding hood: A comprehensive, natural representation for language understanding. In Iwanska, L. and S.C. Shapiro, editors, *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. MIT/AAAI Press.

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

Schubert, Lenhart K. 2005. Some Knowledge Representation and Reasoning Requirements for Self-awareness. In *Proc. AAAI Spring Symposium on Metacognition in Computation*.

Snow, Rion, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of NIPS 17*.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge unifying WordNet and Wikipedia. In *Proceedings of WWW*.