Growing Finely-Discriminating Taxonomies from Seeds of Varying Quality and Size

Tony Veale
School of Computer Science
University College Dublin
Ireland
tony.veale@ucd.ie

Guofu Li
School of Computer Science
University College Dublin
Ireland
guofu.li@ucd.ie

Yanfen Hao
School of Computer Science
University College Dublin
Ireland
yanfen.hao@ucd.ie

Abstract

Concept taxonomies offer a powerful means for organizing knowledge, but this organization must allow for many overlapping and fine-grained perspectives if a general-purpose taxonomy is to reflect concepts as they are actually employed and reasoned about in everyday usage. We present here a means of bootstrapping finely-discriminating taxonomies from a variety of different starting points, or seeds, that are acquired from three different sources: WordNet, ConceptNet and the web at large.

1 Introduction

Taxonomies provide a natural and intuitive means of organizing information, from the biological taxonomies of the Linnaean system to the layout of supermarkets and bookstores to the organizational structure of companies. Taxonomies also provide the structural backbone for ontologies in computer science, from common-sense ontologies like Cyc (Lenat and Guha, 1990) and SUMO (Niles and Pease, 2001) to lexical ontologies like WordNet (Miller *et al.*, 1990). Each of these uses is based on the same root-branch-leaf metaphor: the broadest terms with the widest scope occupy the highest positions of a taxonomy, near the root, while specific terms with the most local concerns are located lower in the hierarchy, nearest the leaves. The more interior nodes that a taxonomy possesses, the finer the conceptual distinctions and the more gradated the similarity judgments it can make (e.g., Budanitsky and Hirst, 2006).

General-purpose computational taxonomies are called upon to perform both coarse-grained and fine-grained judgments. In NLP, for instance, the semantics of "eat" requires just enough knowledge to discriminate foods like tofu and cheese from non-foods like wool and steel, while specific applications in the domain of cooking and recipes (e.g., Hammond's (1986) CHEF) require enough discrimination to know that tofu can be replaced with clotted cheese in many recipes because each is a soft, white and bland food.

So while much depends on the domain of usage, it remains an open question as to how many nodes a good taxonomy should possess. Princeton WordNet, for instance, strives for as many nodes as there are word senses in English, yet it also contains a substantial number of composite nodes that are lexicalized not as single words, but as complex phrases. Print dictionaries intended for human consumption aim for some economy of structure, and typically do not include the meaning of phrases that can be understood as straightforward compositions of the meaning of their parts (Hanks, 2004). But WordNet also serves another purpose, as a lexical knowledge-base for computers, not humans, a context in which concerns about space seem quaint. When space is not a issue, there seems no good reason to exclude nodes from a concept taxonomy merely for being composites of other ideas; the real test of entry is whether a given node adds value to a taxonomy, by increasing its level of internal organization through the systematic dissection of overly broad categories into finer, more intuitive and manageable clusters.

In this paper we describe a means by which finely-discriminating taxonomies can be *grown* from a variety of different knowledge *seeds*. These taxonomies comprise composite categories that can be lexicalized as phrases of the form "ADJ NOUN", such as Sharp-Instrument, which represents the set of all instruments that are typically considered sharp, such as knives, scissors, chisels and can-openers. While WordNet already contains an equivalent category, named Edge-

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 835–842,
Athens, Greece, 30 March – 3 April 2009.

Tool, which it defines with the gloss "any cutting tool with a sharp cutting edge", it provides no structural basis for inferring that any member of this category can be considered *sharp*. For the most part, if two ideas (word senses) belong to the same semantic category X in WordNet, the most we can infer is that both possess the trivial property *X-ness*. Our goal here is to construct taxonomies whose form makes explicit the actual properties that accrue from membership in a category.

Past work on related approaches to taxonomy creation are discussed in section 2, while section 3 describes the different knowledge seeds that serve as the starting point for our bootstrapping process. In section 4 we describe the bootstrapping process in more detail; such processes are prone to noise, so we also discuss how the acquired categorizations are validated and filtered after each bootstrapping cycle. An evaluation of the key ideas is then presented in section 5, to determine which seed yields the highest quality taxonomy once bootstrapping is completed. The paper then concludes with some final remarks in section 6.

2 Related Work

Simple pattern-matching techniques can be surprisingly effective for the extraction of lexico-semantic relations from text when those relations are expressed using relatively stable and unambiguous syntagmatic patterns (Ahlswede and Evens, 1988). For instance, the work of Hearst (1992) typifies this surgical approach to relation extraction, in which a system fishes in a large text for particular word sequences that strongly suggest a semantic relationship such as hypernymy or, in the case of Charniak and Berland (1999), the part-whole relation. Such efforts offer high precision but can exhibit low recall on moderate-sized corpora, and extract just a tiny (but very useful) subset of the semantic content of a text. The *KnowItAll* system of Etzioni *et al.* (2004) employs the same generic patterns as Hearst (e.g., "NPs such as NP1, NP2, …"), and more besides, to extract a whole range of facts that can be exploited for web-based question-answering. Cimiano and Wenderoth (2007) also use a range of Hearst-like patterns to find text sequences in web-text that are indicative of the lexico-semantic properties of words; in particular, these authors use phrases like "to * a new NOUN" and "the purpose of NOUN is to *" to identify the formal (isa), agentive (made by) and telic (used for) roles of nouns.

Snow, Jurafsky and Ng (2004) use supervised learning techniques to acquire those syntagmatic patterns that prove most useful for extracting hypernym relations from text. They train their system using pairs of WordNet terms that exemplify the hypernym relation; these are used to identify specific sentences in corpora that are most likely to express the relation in lexical terms. A binary classifier is then trained on lexico-syntactic features that are extracted from a dependency-structure parse of these sentences. Kashyap *et al.*, (2005) experiment with a bootstrapping approach to growing concept taxonomies in the medical domain. A gold standard taxonomy provides terms that are used to retrieve documents which are then hierarchically clustered; cohesiveness measures are used to yield a taxonomy of terms that can then further drive the retrieval and clustering cycle. Kozareva *et al.* (2008) use a bootstrapping approach that extends the fixed-pattern approach of Hearst (1992) in two intriguing ways. First, they use a doubly-anchored retrieval pattern of the form "$NOUN_{cat}$ such as $NOUN_{example}$ and *" to ground the retrieval relative to a known example of hypernymy, so that any values extracted for the wildcard * are likely to be coordinate terms of $NOUN_{example}$ and even more likely to be good examples of $NOUN_{cat.}$ Secondly, they construct a graph of terms that co-occur within this pattern to determine which terms are supported by others, and by how much. These authors also use two kinds of bootstrapping: the first variation, dubbed *reckless*, uses the candidates extracted from the double-anchored pattern (via *) as exemplars ($NOUN_{example}$) for successive retrieval cycles; the second variation first checks whether a candidate is sufficiently supported to be used as an exemplar in future retrieval cycles.

The approach we describe here is most similar to that of Kozareva *et al*. (2008). We too use a double-anchored pattern, but place the anchors in different places to obtain the query patterns "ADJ_{cat} $NOUN_{cat}$ such as *" and "ADJ_{cat} * such as $NOUN_{example}$". As a result, we obtain a finely-discriminating taxonomy based on categories that are explicitly annotated with the properties (ADJ_{cat}) that they bequeath to their members. These categories have an obvious descriptive and organizational utility, but of a kind that one is unlikely to find in conventional resources like WordNet and Wikipedia. Kozareva *et al*. (2008) test their approach on relatively simple and objective categories like *states*, *countries* (both

closed sets), *singers* and *fish* (both open, the former more so than the latter), but not on complex categories in which members are tied both to a general category, like *food*, and to a stereotypical property, like *sweet* (Veale and Hao, 2007). By validating membership in these complex categories using WordNet-based heuristics, we can hang these categories and members on specific WordNet senses, and thus enrich WordNet with this additional taxonomic structure.

3 Seeds for Taxonomic Growth

A fine-grained taxonomy can be viewed as a set of triples $T_{ijk} = <C_i, D_j, P_k>$, where C_i denotes a child of the parent term P_k that possesses the discriminating property D_j; in effect, each such triple expresses that C_i is a specialization of the complex taxonym D_j-P_k. Thus, the belief that cola is a carbonated-drink is expressed by the triple <*cola, carbonated, drink*>. From this triple we can identify other categorizations of *cola* (such as *treat* and *refreshment*) via the web query "carbonated * such as cola", or we can identify other similarly fizzy drinks via the query "carbonated drinks such as *". So this web-based bootstrapping of fine-grained category hierarchies requires that we already possess a collection of fine-grained distinctions of a relatively high-quality. We now consider three different starting points for this bootstrapping process, as extracted from three different resources: WordNet, ConceptNet and the web at large.

3.1 WordNet

The noun-sense taxonomy of WordNet makes a number of fine-grained distinctions that prove useful in clustering entities into smaller and more natural groupings. For instance, WordNet differentiates *{feline, felid}* into the sub-categories *{true_cat, cat}* and *{big_cat, cat}*, the former serving to group domesticated cats with other cats of a similar size, the latter serving to cluster cats that are larger, wilder and more exotic. However, such fine-grained distinctions are the exception rather than the norm in WordNet, and not one of the 60+ words of the form *Xess* in WordNet that denote a person (such as huntr*ess,* waitr*ess,* Jew*ess*, etc.) express the defining property *female* in explicit taxonomic terms. Nonetheless, the free-text glosses associated with WordNet sense-entries often do state the kind of distinctions we would wish to find expressed as explicit taxonyms. A shallow parse of these glosses thus yields a sizable number of fine-grained distinctions, such as <*lioness, female, lion*>, <*espresso, strong, coffee*> and both <*messiah, awaited, king*> and <*messiah, expected, deliverer*>.

3.2 ConceptNet

Despite its taxonomic organization, WordNet owes much to the centralized and authority-preserving craft of traditional lexicography. ConceptNet (Liu and Singh, 2004), in contrast, is a far less authoritative knowledge-source, one that owes more to the workings of the WWW than to conventional print dictionaries. Comprising factoids culled from the template-structured contributions of thousands of web users, ConceptNet expresses many relationships that accurately reflect a public, common-sense view on a given topic (from vampires to dentists) and many more that are simply bizarre or ill-formed. Looking to the relation that interests us here, the IsA relation, ConceptNet tells us that an *espresso* is a *strong coffee* (correctly, like WordNet) but that a *bagel* is a *Jewish word* (confusing *use* with *mention*). Likewise, we find that *expressionism* is an *artistic style* (correct, though WordNet deems it an *artistic movement*) but that an *explosion* is a *suicide attack* (confusing formal and telic roles). Since we cannot trust the content of ConceptNet directly, lest we bootstrap from a highly unreliable starting point, we use WordNet as a simple filter. While the concise form of ConceptNet contains over 30,000 IsA propositions, we consider as our seed collection only those that define a noun concept (such as "espresso") in terms of a binary compound (e.g., "strong coffee") where the head of the latter (e.g., "coffee") denotes a WordNet hypernym of some sense of the former. This yields triples such as <*Wyoming, great, state*>, <*wreck, serious, accident*> and <*wolf, wild, animal*>.

3.3 Web-derived Stereotypes

Veale and Hao (2007) also use the observations of web-users to acquire common perceptions of oft-mentioned ideas, but do so by harvesting simile expressions of the form "as ADJ as a NOUN" directly from the web. Their approach hinges on the fact that similes exploit stereotypes to draw out the salient properties of a target, thereby allowing rich descriptions of those stereotypes to be easily acquired, e.g., that snowflakes are pure and unique, acrobats are agile and nimble, knifes are sharp and dangerous, viruses are malicious and infectious, and so on. However, because they find that almost 15% of their web-harvested sim-

iles are ironic (e.g., "as subtle as a rock", "as bulletproof as a sponge-cake", etc.), they filter irony from these associations by hand, to yield a sizable database of stereotypical attributions that describes over 6000 noun concepts in terms of over 2000 adjectival properties. However, because Veale and Hao's data directly maps stereotypical properties to simile vehicles, it does not provide a parent category for these vehicles. Thus, the seed triples derived from this data are only partially instantiated; for instance, we obtain <*surgeon, skilful, ?*>, <*virus, malicious, ?*> and <*dog, loyal, ?*>. This does not prove to be a serious impediment, however, as the missing field of each triple is quickly identified during the first cycle of bootstrapping.

3.4 Overview of Seed Resources

Neither of these three seeds is an entirely useful knowledge-base in its own right. The WordNet-based seed is clearly a representation of convenience, since it contains only those properties that can be acquired from the glosses that happen to be amenable to a simple shallow-parse. The ConceptNet seed is likewise a small collection of low-hanging fruit, made smaller still by the use of WordNet as a coarse but very necessary noise-filter. And while the simile-derived distinctions obtained from Veale and Hao paint a richly detailed picture of the most frequent objects of comparison, this seed offers no coverage for the majority of concepts that are insufficiently noteworthy to be found in web similes. A quantitative comparison of all three seeds is provided in Table 1 below.

	WordNet	ConceptNet	Simile
# terms in total	12,227	1,133	6512
# triples in total	51,314	1808	16,688
# triples per term	4.12	1.6	2.56
# features	2305	550	1172

Table 1: The size of seed collections yielded from different sources.

We can see that WordNet-derived seed is clearly the largest and apparently the most comprehensive knowledge-source of the three: it contains the most terms (concepts), the most features (discriminating properties of those concepts), and the most triples (which situate those concepts in parent categories that are further specialized by these discriminating features). But size is only weakly suggestive of quality, and as we shall see in the next section, even such dramatic differences in scale can disappear after several cycles of bootstrapping. In section 5 we will then consider which of these seeds yields the highest quality taxonomies after bootstrapping has been applied.

4 Bootstrapping from Seeds

The seeds of the previous section each represent a different starting collection of triples. It is the goal of the bootstrapping process to grow these collections of triples, to capture more of the terms – and more of the distinctions – that a taxonomy is expected to know about. The expansion set of a triple T_{ijk} = <C_i, D_j, P_k> is the set of triples that can be acquired from the web using the following query expansions (* is a search wildcard):

1. "D_j * such as C_i"
2. "D_j P_k such as *"

In the first query, a noun is sought to yield another categorization of C_i, while in the second, other members of the fine-grained category D_j-P_k are sought to accompany C_i. In parsing the text snippets returned by these queries, we also exploit text sequences that match the following patterns:

3. "* and D_j P_k such as *"
4. "* and D_j * such as C_i"

These last two patterns allow us to learn new discriminating features by noting how these discriminators are combined to reinforce each other in some ad-hoc category formulations. For instance, the phrase "cold and refreshing beverages such as lemonade" allows us to acquire the triples <*lemonade, cold, beverage*> and <*lemonade, refreshing, beverage*>. This pattern is necessary if the bootstrapping process is to expand beyond the limited vocabulary of discriminating features (D_j) found in the original seed collections of triples.

We denote the mapping from a triple T to the set of additional triples that can be acquired from the web using the above queries/patterns as *expand(T')*. We currently implement this function using the Google search API. Our experiences with each query suggest that 200 snippets is a good search range for the first query, while 50 is usually more than adequate for the second.

We can now denote the knowledge that is acquired when starting from a given seed collection S after t cycles of bootstrapping as K_t^S. Thus,

$$K_0^S = S$$
$$K_1^S = K_0^S \cup \{T \mid T' \in S \wedge T \in expand(T')\}$$
$$K_{t+1}^S = K_t^S \cup \{T \mid T' \in K_t^S \wedge T \in expand(T')\}$$

Web queries, and the small snippets of text that they return, offer just a keyhole view of language as it is used in real documents. Unsurprisingly, the new triples acquired from the web via *expand(T')* are likely to be very noisy indeed. Following Kozareva *et al.* (2008), we can either indulge in *reckless bootstrapping*, which ignores the question of noise until all bootstrapping is finished, or we can apply a noise filter after each incremental step. The latter approach has the additional advantage of keeping the search-space as small as possible, which is a major consideration when bootstrapping from sizable seeds. We use a simple WordNet-based filter called *near-miss*: a new triple <C_i, D_j, P_k> is accepted if WordNet contains a sense of C_i that is a descendant of some sense of P_k (a hit), or a sense of C_i that is a descendant of the direct hypernym of some sense of P_k (a near-miss). This allows the bootstrapping process to acquire structures that are not simply a decorated version of the basic WordNet taxonomy, but to acquire hierarchical relations whose undifferentiated forms are not in WordNet (yet are largely compatible with WordNet). This non-reckless bootstrapping process can be expressed as follows:

$$K_{t+1}^S = K_t^S \cup \left\{T \,\middle|\, T' \in K_t^S \wedge T \in filter_{near-miss}(expand(T'))\right\}$$

Figure 1 and figure 2 below illustrate the rate of growth of triple-sets from each of our three seeds.

Referring again to table 1, we note that while the ConceptNet collection is by far the smallest of the three seeds – more that 7 times smaller than the simile-derived seed, and almost 40 times smaller than the WordNet seed – this difference is size shrinks considerably over the course of five bootstrapping cycles. The WordNet *near-miss* filter ensures that the large body of triples grown from each seed are broadly sound, and that we are not simply generating comparable quantities of nonsense in each case.

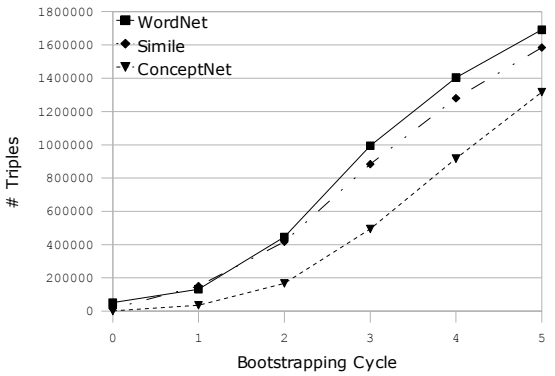

Figure 1: Growth in the number of acquired triples, over 5 cycles of bootstrapping from different seeds.

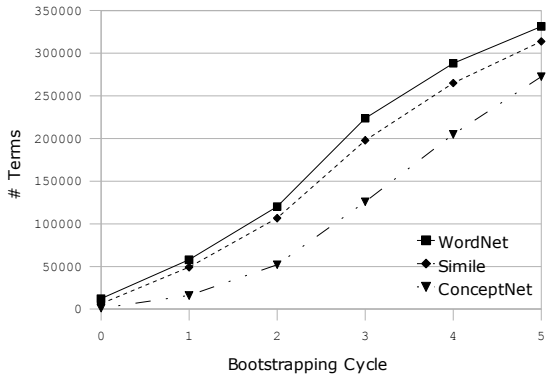

Figure 2: Growth in the number of terms described by the acquired triples, over 5 cycles of bootstrapping from different seeds.

4.1 An Example

Consider *cola*, for which the simile seed has one triple: <*cola, refreshing, beverage*>. After a single cycle of bootstrapping, we find that *cola* can now be described as an *effervescent beverage*, a *sweet beverage*, a *nonalcoholic beverage* and more. After a second cycle, we find it described as a *sugary food*, a *fizzy drink* and a *dark mixer*. After a third cycle, it is found to be a *sensitive beverage*, an *everyday beverage* and a *common drink*. After a fourth cycle, it is also found to be an *irritating food* and an *unhealthy drink*. After the fifth cycle, it is found to be a *stimulating drink*, a *toxic food* and a *corrosive substance*. In all, the single *cola* triple in the simile seed yields 14 triples after 1 cycle, 43 triples after 2 cycles, 72 after 3 cycles, 93 after 4 cycles, and 102 after 5 cycles. During these bootstrapping cycles, the description *refreshing beverage* additionally becomes associated with the terms *champagne*, *lemonade* and *beer*.

5 Empirical Evaluation

The WordNet *near-miss* filter thus ensures that the parent field (P_k) of every triple contains a value that is sensible for the given child concept (C_i), but does not ensure that the discriminating property (D_j) in each triple is equally sensible and apropos. To see whether the bootstrapping process is simply padding the seed taxonomy with large quantities of noise, or whether the acquired D_j values do indeed mark out the implicit essence of the C_i terms they describe, we need an evaluation framework that can quantify the ontological usefulness of these D_j values. For this, we use the experimental setup of Almuhareb and Poesio (2005), who use information extraction from the web to acquire attribute values for different terms/concepts, and who then compare the taxonomy that can be induced by clustering these values with the taxonomic backbone of WordNet.

Almuhareb and Poesio first created a balanced set of 402 nouns from 21 different semantic classes in WordNet. They then acquired attested attribute values for these nouns (such as *hot* for coffee, *red* for car, etc.) using the query "*(a|an|the)* * C_i *(is|was)*" to find corresponding D_j values for each C_i. Unlike our work, these authors did *not* seek to acquire hypernyms for each C_i during this search, and did not try to link the acquired attribute values to a particular branching point (P_k) in the taxonomy (they did, however, seek matching attributes for these values, such as *Temperature* for *hot*, but that aspect is not relevant here). They acquired 94,989 attribute values in all for the 402 test nouns. These values were then used as features of the corresponding nouns in a clustering experiment, using the CLUTO system of Karypis (2002). By using attribute values as a basis for partitioning the set of 402 nouns into 21 different categories, Almuhareb and Poesio attempted to reconstruct the original 21 WordNet categories from which the nouns were drawn. The more accurate the match to the original WordNet clustering, the more these attribute values can be seen (and used) as a representation of conceptual structure. In their first attempt, they achieved just a 56.7% clustering accuracy against the original human-assigned categories of WordNet. But after using a noise-filter to remove almost half of the web-harvested attribute values, they achieve a higher cluster accuracy of 62.7%. More specifically, Poesio and Almuhareb achieve a cluster purity of 0.627 and a cluster entropy of 0.338 using 51,345 features to describe and cluster the 402 nouns.[1]

We replicate the above experiments using the same 402 nouns, and assess the clustering accuracy (again using WordNet as a gold-standard) after each bootstrapping cycle. Recall that we use only the D_j fields of each triple as features for the clustering process, so the comparison with the WordNet gold-standard is still a fair one. Once again, the goal is to determine how much like the human-crafted WordNet taxonomy is the taxonomy that is clustered automatically from the discriminating words D_j only. The clustering accuracy for all three seeds are shown in Tables 2, 3 and 4.

Cycle	E	P	# Features	Coverage
1st	.327	.629	907	66%
2nd	.253	.712	1,482	77%
3rd	.272	.717	2,114	82%
4th	.312	.640	2,473	83%
5th	.289	.684	2,752	83%

Table 2: Clustering accuracy using the *WordNet* seed collection (E denotes Entropy and P stands for Purity)

Cycle	E	P	# Features	Coverage
1st	.115	.842	363	41%
2nd	.255	.724	787	59%
3rd	.286	.694	1,362	74%
4th	.279	.694	1,853	79%
5th	.299	.673	2,274	82%

Table 3: Clustering accuracy using the *ConceptNet* seed collection

Cycle	E	P	# Features	Coverage
1st	.254	.716	837	59%
2nd	.280	.712	1,338	73%
3rd	.289	.693	1,944	79%
4th	.313	.660	2,312	82%
5th	.157	.843	2,614	82%

Table 4: Clustering accuracy using the *Simile* seed collection

The test-set of 402 nouns contains some low-frequency words, such as *casuarina, cinchona, dodecahedron*, and *concavity*, and Almuhareb and

[1] We use cluster purity as a reflection of clustering accuracy. We express accuracy as a percentage; hence a purity of 0.627 is seen as an accuracy of 62.7%.

Poesio note that one third of their data-set has a low-frequency of between 5-100 occurrences in the British National Corpus. Looking to the coverage column of each table, we thus see that there are words in the Poesio and Almuhareb data set for which no triples can be acquired in 5 cycles of bootstrapping. Interestingly, though each seed is quite different in origin and size (see again Table 1), all reach similar levels of coverage (~82%) after 5 bootstrapping cycles. Test nouns for which all three seeds fail to reach a description include *yesteryear, nonce* (very rare), *salient* (more typically an adjective), *jag, droop, fluting, fete, throb, poundage, stinging, rouble, rupee, riel, drachma, escudo, dinar, dirham, lira, dispensation*, *hoard*, *airstream* (not typically a solid compound), *riverside* and *curling*. Figures 3 and 4 summarize the key findings in the above tables: while bootstrapping from all three seeds converges to the same level of coverage, the simile seed clearly produces the highest quality taxonomy.

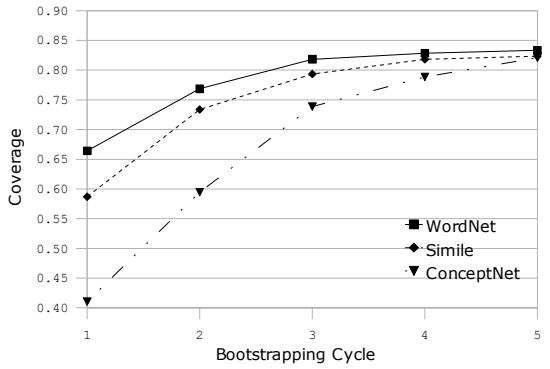

Figure 3: Growth in the coverage from different seed sources.

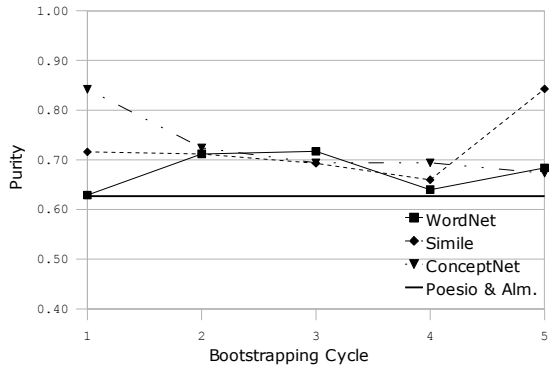

Figure 4: Divergence in the clustering Purity achieved using different seed sources. The results of Poesio and Almuhareb are shown as the straight line: y = 0.627.

Both the WordNet and ConceptNet seeds achieve comparable accuracies of 68% and 67% respectively after 5 cycles of bootstrapping, which compares well with the accuracy of 62.7% achieved by Poesio and Almuhareb. However, the simile seed clearly yields the best accuracy of 84.3%, which also exceeds the accuracy of 66.4% achieved by Poesio and Almuhareb when using both values *and* attributes (such as *Temperature*, *Color*, etc.) for clustering, or the accuracy of 70.9% they achieve when using attributes alone. Furthermore, bootstrapping from the simile seed yields higher cluster accuracy on the 402-noun data-set than Veale and Hao (2008) themselves achieve with their simile data on the same test-set (69.85%).

But most striking of all is the concision of the representations that are acquired using bootstrapping. The simile seed yields a high cluster accuracy using a pool of just 2,614 fine discriminators, while Poesio and Almuhareb use 51,345 features even after their feature-set has been filtered for noise. Though starting from different initial scales, each seed converges toward a feature-set that is roughly twenty times smaller than that used by Poesio and Almuhareb.

6 Conclusions

These experiments reveal that seed knowledge of different authoritativeness, quality and size will tend to converge toward roughly the same number of finely discriminating properties and toward much the same coverage after 5 or so cycles of bootstrapping. Nonetheless, quality wins out, and the simile-derived seed knowledge shows itself to be a clearly superior basis for reasoning about the structure and organization of conceptual categories. Bootstrapping from the simile seed yields a slightly smaller set of discriminating features than bootstrapping from the WordNet seed, one that is many times smaller than the Poesio and Almuhareb feature set. What matters is that they are the right features to discriminate with.

There appears to be a number of reasons for this significant difference in quality. For one, Veale and Hao (2007) show that similes express highly stereotypical beliefs that strongly influence the affective disposition of a term/concept; negatively perceived concepts are commonly used to exemplify negative properties in similes, while positively perceived concepts are widely used to exemplify positive properties. Veale and Hao (2008) go on to argue that similes offer a very concise snapshot of those widely-held beliefs that are the cornerstone of everyday reason-

ing, and which should thus be the corner-stone of any general-purpose taxonomy. In addition, beliefs expressed via the "as D_j as C_i" form of similes appear to lend themselves to re-expression via the "D_j P_k such as C_i" form; in each case, a concept C_i is held up as an exemplar of a salient property D_j. Since the "such as" bootstrapping pattern seeks out expressions of prototypicality on the web, a simile-derived seed set is likely the best starting point for this search.

All three seeds appear to suffer the same coverage limitations, topping out at about 82% of the words in the Poesio and Almuhareb data-set. Indeed, after 5 bootstrapping cycles, all three seeds give rise to taxonomies that overlap on 328 words from the 402-noun test-set, accounting for 81.59% of the test-set. In effect then, bootstrapping stumbles over the same core of hard words in each case, no matter the seed that is used. As such, the problem of coverage lies not in the seed collection, but in the queries used to perform the bootstrapping. The same coverage limitations will thus apply to other bootstrapping approaches to knowledge acquisition, such as Kozareva *et al.* (2008), which rely on much the same stock patterns. So while bootstrapping may not be a general solution for acquiring all aspects of a general-purpose taxonomy, it is clearly useful in acquiring large swathes of such a taxonomy if given a sufficiently high-quality seed to start from.

References

Ahlswede, T. and Evans, M. (1988). Parsing vs. Text Processing in the analysis of dictionary definitions. *In Proc. of the 26th* Annual Meeting of the ACL, pp 217-224.

Almuhareb, A. and Poesio, M. (2005). Concept Learning and Categorization from the Web. In *Proc. of the annual meeting of the Cognitive* Science *Society*, Italy, July.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics,* 32(1):13-47.

Cimiano, P. and Wenderoth, J. (2007). Automatic Acquisition of Ranked Qualia Structures from the Web. *In Proc. of the 45th Annual Meeting of the ACL*, pp 888-895.

Charniak, E. and Berland, M. (1999). Finding parts in very large corpora. *In Proc. of the 37th Annual Meeting of the ACL*, pp 57–64.

Etzioni, O., Kok, S., Soderland, S., Cafarella, M., Popescu, A-M., Weld, D., Downey, D., Shaked, T. and Yates, A. (2004). Web-scale information extraction in KnowItAll (preliminary results). *In Proc. of the 13th WWW Conference*, pp 100–109.

Hammond, K. J. (1986). CHEF : A Model of Case--based Planning. *In Proc. of the 5th National Conference on Artificial Intelligence*, pp 267--271, Philadelphia, Pennsylvania. American Association for Artificial Intelligence.

Hanks, P. (2004). WordNet: What is to be done? In *Proc. of GWC'2004, the 2nd Global WordNet conference*, Masaryk University, Brno.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. *In Proc. of the 14th Int. Conf. on Computational Linguistics*, pp 539–545.

Kashyap, V. Ramakrishnan, C. and Sheth, T. A. (2005). TaxaMiner: an experimentation framework for automated taxonomy bootstrapping. *Int. Journal of Web and Grid Services* 1(2), pp 240-266.

Karypis, G. (2002). CLUTO: A clustering toolkit. *Technical Report 02-017*, University of Minnesota. http://www-users.cs.umn.edu/~karypis/cluto/.

Kozareva, Z., Riloff, E. and Hovy, E. (2008). Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. *In Proc. of the 46th Annual Meeting of the ACL.*

Lenat, D. B. and Guha, R. V. (1990). Building large knowledge-based systems: representation and inference in the Cyc project. NY: Addison-Wesley.

Liu, H. and Singh, P. (2004), ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal*, 22(4):211-226.

Miller, G., Beckwith,R., Fellbaum, C., Gross, D. and Miller, K.J. (1990). Introduction to WordNet: an on-line lexical database. Int. Journal of Lexicography, 3(4):235 – 244.

Niles, I. and Pease, A. (2001). Toward a standard upper ontology. *In Proc. of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*.

Snow, R., Jurafsky, D. and Ng, A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems* 17.

Veale, T. and Hao, Y. (2007). Making Lexical Ontologies Functional and Context-Sensitive. *In Proc. of the 45th Annual Meeting of the ACL*, pp 57–64.

Veale, T. and Hao, Y. (2008). A Fluid Knowledge Representation for Understanding and Generating Creative Metaphors. *In Proc. of Coling 2008, The 22nd International Conference on Computational Linguistics*, Manchester.