Corpus-Driven Contextualized Categorization

Tony Veale and **Yanfen Hao**[1]

Abstract. Ontologies strive to offer a interconnected, hierarchical systems of categories to guide our actions in a complex world. But the boundaries of these categories are highly context-dependent, and what constitutes a prototypical category member in one context may be atypical or unrepresentative in another. In this paper we outline a dynamic, trainable, bottom-up view of category structure based on context-sensitive corpus analysis. By learning from corpora about how people creatively actually use categories in different contexts, we can train our ontologies to creatively adapt themselves to these contexts.

1 INTRODUCTION

An ontology is a system of inter-connected categories that collectively provide a structured representation of a given domain. As such, an ontology serves as the conceptual bedrock against which domain meanings are constructed, manipulated and interpreted. However, this fundamental role of the ontology should not blind us to the fact that much of what an ontology attempts to model, via its category structure, is not static but dynamic, making the use of these categories highly sensitive to context. Consider that many categories in a language-oriented ontology, like Genius, Fool, Hero, Villain, Expert, Hunter, and so on, possess subjective membership criteria that change from user to user, and from context to context. Are politicians fools, villains or schemers? Are firemen heroes or workmen? Are scientists experts or geniuses?

Since top-down definitions of membership criteria will always seem brittle or inadequate in some contexts, it seems best to allow contexts to define their own criteria, bottom-up. In other words, we need to establish a contextual ontology [10] based category structure, which not only preserves the common view of concepts, but also keeps the local perspective of domains. For language-oriented ontologies, like WordNet [6] (a flawed, lightweight ontology to be sure, but an ontology none the less), HowNet [1] and, to some extent, Cyc [5], the context of usage can conveniently be captured via a large corpus of representative texts. A corpus-based approach to determining category membership allows us to structure the middle and lower layers of an ontology according to how words and concepts are actually used in a particular domain. In short, a corpus-based approach supports an extremely flexible, non-classical view of category structure, one that views category membership as a graded rather than binary notion [4], and one in which concepts can fluidly move (via metaphor) from one category to another [2]. In this current work, we use the ability to support metaphoric reasoning as the yardstick against which ontological flexibility should be measured.

Of course, this fluidity does not sit well with conventional perspectives on ontological structure, as represented by the ontologies of [1,5,6]. In this paper we look at one conventional ontology, the HowNet system of [1], which is a large-scale bilingual lexical ontology for words and their meanings in both Chinese and English. In many respects, HowNet is similar to the WordNet lexical ontology for English [6], though in contrast to WordNet, HowNet provides an explicit, if sparse, propositional semantics for each of the word-concepts it defines. Complementing this frame-like semantics, in which concepts are defined in terms of actions, case-roles and fillers, is a taxonomic backbone that seems rather impoverished when compared to that of WordNet. HowNet is essentially an ontology of ”Being” rather than an ontology of ”Doing” which is to say that it defines concepts according to conventional kinds like *human*, *animal*, *tool* and so on - rather than according to how specific concepts actually behave in context. However, we describe in section 2 how HowNet's propositional semantics can be used to automatically derive an ontology of ”Doing” to replace HowNet's rather shallow taxonomy of conventional categories [8]. Once in place, we demonstrate how this new system of derived categories can be made contextually sensitive by defining their membership criteria in statistical, corpus-based terms, to create a fluid system of membership akin to the Slipnets of Hofstadter [3]. Once sensitized in this way, the ontology can be moved with ease from one context to another simply by replacing the underlying corpus.

2 ONTOLOGIES OF ”BEING” AND ”DOING”

HowNet and WordNet each reflect a different view of semantic organization. WordNet [7] is *differential* in nature: rather than attempting to express the meaning of a word explicitly, WordNet instead differentiates words with different meanings by placing them in different synonym sets, or *synsets*, and further differentiates these synsets from one another by assigning them to different positions of a taxonomy. In contrast, HowNet is *constructive* in nature. It does not provide a human-oriented textual gloss for each lexical concept, but instead composes sememes from a less discriminating taxonomy to provide a semantic representation for each word sense. For example, HowNet defines the lexical concept *surgeon*|医生 as follows:

(1)*surgeon*|医生 {*human*|人 :*HostOf*={*Occupation*|职位}
domain={*medical*|医}}, {*doctor*|医治:*agent*={∼}}}

which can be glossed thus: ”a surgeon is a human, with an occupation in the medical domain, who acts as an agent of a doctoring activity” (the {∼} here serves to indicate the placement of the concept within its associated propositional structure). We see a similar structure employed by HowNet for the lexical concept *repairman*|修理工:

[1] School of Computer Science and Informatics, University College Dublin, Ireland, email: {tony.veale, yanfen.hao}@ucd.ie

(2)*repairman*|修理工 {*human*|人:*HostOf*={*Occupation*|职位},
{*repair*|修理:*agent*={∼}}}

Note that the impoverished nature of HowNet's taxonomy means that over 3000 different concepts are forced to share the immediate hypernym *human*|人. However, *human*|人 merely states, very generally, what a repairman is, rather than what a repairman does. Fortunately, HowNet also organizes its verb entries taxonomically, and so we find the verbs *doctor*|医治 and *repair*|修理 organized under the hypernym *resume*|恢复 (the logic being, one supposes, that "doctoring" and "repairing" both involve a resumption of an earlier, better state). This similarity of verbs, combined with an identicality of case-roles (both surgeon and repairman are agents of their respective activities), allows us to abstract out a new taxonomy, based on the behaviour rather than the general type of these entities.

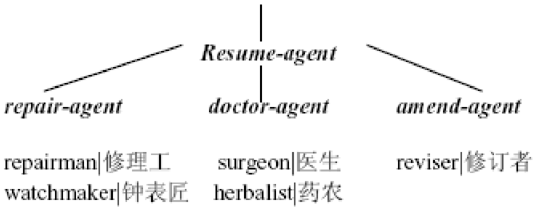

Figure 1. A new 3-level abstraction hierarchy derived from verb/role combinations.

Figure 1 illustrates the creation of such a taxonomy, whose categories represent a yoking of verbs to specific case-roles, such as *repair-agent* and *amend-agent*, and whose category members are those HowNet concepts defined using these verbs and roles. The category-hopping nature of metaphor is now rather easily construed as a combination of generalization and re-specialization operations, in which one moves from one category to another by first passing through a common super-category like *resume-agent*. Thus, a surgeon can be seen as a repairman or a watchmaker, while a reviser of texts (an editor) can sometimes be seen as a surgeon. These metaphors make sense not because each is a human, but because each restores a better state.

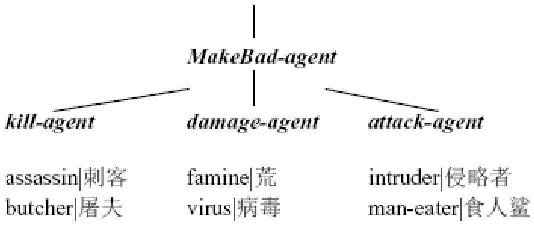

Figure 2. Newly derived HowNet categories may contain a diverse range of concepts.

Of course, this Aristotelian view of metaphor as an abstract "carrying-over" (the etymological origin of the word "metaphor") can only be valid if concepts are ontologized by what they do, rather than by what they are (as is typically the case, in both WordNet and HowNet, and even Cyc [6]). Otherwise, metaphor could never operate between semantically distant concepts, which it plainly does. For instance, figure 2 illustrates the derived taxonomy for HowNet concepts that are defined as agents of the verbs "kill", "damage" and "attack", each a specialization of the abstract verb *MakeBad* in HowNet. We see in this taxonomy the potential for famines to be metaphorically viewed as butchers and assassins, and for viruses to be seen as deadly intruders, or even man-eaters.

3 DERIVING FLUID CATEGORY STRUCTURES

An ontology of "doing" begs a number of obvious questions about the nature of categorization. For instance, is every concept that kills an equally representative member of the category *kill-agent*? Is movement always allowed between any two categories that share a common abstraction like *MakeBad-agent*, or is movement limited to certain members only, and in certain directions? When a concept moves from its conventional category to another, how is its degree of membership in this new category to be assessed? In this section we address this key issue of obtaining fluid category structure.

There are two major approaches in the community of automatic acquisition of taxonomies. One approach is based on the distributional hypothesis made by Harris[11], in which he believes that word terms are similar if they have similar linguistic contexts. For instance, Hindle[12] clusters nouns according to their contextual attributes, such as the co-occurrence of nouns with verbs as subjects or objects. Steffen Staab[13] also extracts context information (verb/subject dependencies, verb/object dependencies, e.g.) about a certain term from corpus and applies a Formal Concept Analysis to generate a lattice that is finally transformed into a partial order closer to a concept hierarchy. Another major approach is on the basis of investigating the ontological relations such as is-a relation, part-of relation, e.g. via the corpus. Hearst[14] is a representative of this field. However, it seems that these approaches still result in binary and static taxonomies because they all apply the threshold to the category or the concept architecture to determine whether or not a word concept belongs to it. In our approach, we also follow Harris[11]'s distributional hypothesis to investigate the contextual attributes, particularly, the behavior of nouns. The difference is that we apply Lakoff[4]'s category theory to assign the graded membership to nouns within a category rather than simply grouping them into classes according to their contextual attributes or ontological relations.

Following Lakoff [4], every category will possess a prototype, a member that is highly representative of the category as a whole. Such prototypes are often lexicalized in simple terms; for instance, "killer" will be a highly representative of *kill-agent*, while the Chinese translation "杀手" is a composition of "killing" (杀) and "expert" (手). However, many categories like *damage-agent* have no obvious lexicalized prototype, so we need a more generic means of identifying the prototypical member of a category. Lakoff [4] suggests that the prototype will occupy a central position in the category's structure, with other members organized in a radial fashion, at a distance from the centre that is inversely proportional to their similarity to the prototype. If we assume that the prototype will be that member that is most evocative of a category, we should first measure the evocation strength of each concept for a given category. This can be done by determining the frequency of occurrence of each concept within the category, and this, in turn, can be estimated by looking to a large corpus to see how each concept is actually employed by language users. Once the most evocative example is found for each category, membership scores can be assigned based on the strength of evocation. The corpus we use must be large, and while reasonably authoritative it must use words both literally and figuratively. For reasons outlined in section 5, we use here as our corpus the complete text of the open-source encyclopaedia Wikipedia [9].

Thus, to estimate the membership level of the word-concept *butcher*|*屠夫* in the category *kill-agent*, we first determine the corpus-frequency of the phrase ”butcher who kills/killed”. In general, for estimating the membership of the concept C in the category *V-agent*, we use the query form ”C who|which|that V”; for categories of the form *V-instrument*, we use the query ”V with C”, and so on. Of course, some verbs are more vague than others, and can have much higher corpus frequencies. We therefore need to normalize raw corpus-frequencies to obtain a truer picture of evocation power. If f_{raw}(V-role:C) denotes the corpus frequency of concept C when considered as a member of the category *V-role*, where V is a verb like ”kill” and role is one of *agent*, *instrument*, etc., then the adjusted frequency, a measure of true evocation, is estimated by:

$$f_{adj}(\text{V-role:C}) = ln(f_{raw}(\text{V-role:C})) \times ln(\sum_{x} f_{raw}(\text{V-role:x}))^{-1} \quad (1)$$

Now, the prototype will be that member of a category with the strongest evocation:

$$Prototype(\text{V-role}) = max_c(f_{adj}(\text{V-role:C})) \quad (2)$$

The degree of membership of C in the category V-role is relative to the prototype:

$$Membership(\text{V-role:C}) = f_{adj}(\text{V-role:C}) \times f_{adj}(\text{V-role:}prototype(\text{V-role}))^{-1} \quad (3)$$

This ensures that the prototypical member has a membership score of 1, while all other members of a category will have a score in the range 0... 1. A concept can metaphorically be moved from a category in which it is conventionally a member to any other category in which it is considered to have a non-zero membership score.

4 CLUSTERS AND GROUP-TERMS

For ontological purposes, a category is essentially a cluster of concepts that allows one to conveniently infer similarity – the possession of common properties and shared behaviour – from the simple act of co-categorization. That these clusters often have a heterogeneous roster of members (e.g., as illustrated in Figure 2) is testament both to the prevalence of metaphor and to the necessity of viewing ontological categories as categories of ”doing” rather than of ”being”. Of course, the converse is also true: we can infer the contextual behaviour of a concept from how that concept is explicitly clustered with others. And one common way of signalling the appropriate cluster for a concept is through an evocative group word, like ”army”, ”mob”, ”tribe” or ”coven”. For instance, when one uses the phrase ”an army of robots”, one is conveying a soldier-like perspective on the concept Robot, signalling that in this context, Robot should be viewed more as a attacking agent than as a utensil.

Group terms like ”army”, ”family” and ”swarm” are highly suggestive of particular behaviours. For instance, the corpus techniques of section 3 reveal that, in the context of Wikipedia, a ”swarm” has two dominant behaviours, *biting* and *attacking*, while an ”army” has three, *defeating*, *fighting* and *attacking*. To use the phrase ”swarm of X” or ”army of X” is to suggest that X also exhibits these behaviours, and furthermore, that X is similar in behaviour to other concepts that comfortably fit these templates. This intuition is easily contextualized, since the relative frequency of these phrases in a context's corpus will reveal the extent to which different concepts belong to different group-based categories.

As a corpus, Wikipedia is biased toward popular culture and genres such as science fiction. This lack of neutrality makes the Wikipedia corpus an excellent example of a context, more so than traditional language corpora. Consider the population of the category *Army-member* as derived from Wikipedia:

mercenary(238), clone(132), soldier(122), volunteer(72), monster(70), robot(63), minion(60), warrior(60), frog(58), knight(50), slave(48), demon(46), clansman(46), monkey(46), crusader(44), gladiator(38), ant(37), lawyer(32), contributor(28), mutant(27), ...

Note the prominent presence of the genre elements ”clone”, ”robot” and ”minion”, as well as examples like ”lawyer” for which ”army” has a metaphoric meaning. This grouping suggests that lawyers may be seen, alternately, as mercenaries, warriors and even clones, while the extent to which these comparisons are apt in a particular context is a function of how many different groups can contextually claim both as members. For instance, ”lawyer” and ”warrior” are used with seven different group terms in the Wikipedia corpus – *society, family, cadre, team, army, class and squad,* while ”lawyer” and ”mercenary” share just three groupings – *team, army, squad*. Interestingly, the most common group term for ”lawyer” in Wikipedia is ”huddle” (the phrase ”huddle of lawyers” occurs 64 times, twice as often as ”army of lawyers”), which suggests that, in this context, lawyers are more likely to be categorized as players than warriors, mercenaries, clones or robots

5 PRELIMINARY EMPIRICAL EVALUATION

The choice of corpus is clearly key to the quality of category-membership statistics that can be derived using the methods of sections 3 and 4. This corpus must be large, it must be representative of language use in general, and it should offer a means of search that is robust in the face of noise. At first blush, then, the world-wide-web seems an ideal candidate: in size it is unmatched, and various APIs are available to access powerful search engines like Google. Unfortunately, such APIs rarely provide enough control over the query or the archive to ensure that noise can be eliminated, since these engines typically perform their own stemming and stop-word elimination, putting truly strict matching beyond our reach. This means that common noun-noun collocations, like ”fossil record” and ”share issue”, are easily confused for infrequent or nonsensical noun-verb collocations like ”fossils that record” and ”shares that issue”.

To ensure strict matching with controlled morphology, we require a local text corpus that we can index and search directly, and even subject to part-of-speech tagging. For this reason we choose the collected text of the open-source encyclopaedia Wikipedia [9], which is available to download in XML form. Wikipedia has several obvious benefits as a text corpus: each document is explicitly tagged with a subject-label, since each article defines a specific headword; documents exist in a rich web of interconnections; and documents strive to be authoritative on their subjects. Consider the range of subjects that are found in Wikipedia for the verb ”to infect” (with frequencies shown in parentheses):

virus(46), worm(12), retrovirus(7), strain(6), disease(6), bureaucrat(6), poison(4), ally(4), fungus(4), dust(3), smut(2), bacterium(2), physiologist(2), blood(2), plague(2), war(2), substance(2), germ(1), application(1), species(1)

Now consider the range of verbs that can be used with the subject ”virus”:

infect(46), attack(11), kill(7), jump(6), eat(4), drive(3), produce(3), destroy(3), spread(3), transform(3), escape(2), steal(1), prove(1), carry(1), freeze(1), arrive(1), control(1)

We see from this snapshot that Wikipedia contains enough diversity to capture the dominant application of each verb, and the dominant behaviour of each subject noun. Furthermore, Wikipedia contains enough diversity to reveal creative uses of these nouns and verbs; this snapshot reveals, for instance, that ”smut” can ”infect” (2 uses) and that a ”virus” can ”eat”, ”escape” and even ”steal”.

One can ask how well these corpus-derived category structures compare with the hand-crafted category structures of HowNet, since one can reasonably expect human-assigned category memberships to be a gold standard for this task. We find that in 69% of cases, the HowNet-assigned category for a given word-concept is also the dominant corpus-derived category, and that in 76% of cases, a word-concept has a statistical membership in the HowNet-assigned category that is greater than the median membership score for that category.

In fact, these results suggest that HowNet is far from being a gold-standard for category membership. In many cases, the HowNet category name is either poorly named or is dangerously misleading. For instance, the primary sense of the verb ”doctor” in English is not ”heal” but ”fiddle” (as in ”to doctor one’s r*é*sum*é*”). Likewise, HowNet assigns the name ”resume” to the super-category of ”repair” and ”doctor”, when the verb ”restore” is more appropriate in English. In many other cases, the HowNet assigned category is only one of several that seem intuitively appropriate. For instance, the word ”knight” is assigned the dominant category protect-agent (based on 12 occurrences of the pattern ”knight who protects”) while HowNet assigns it to the category defend-agent (which is the second-most popular corpus assignment, based on 10 occurrences of ”knight who defends”). Viewed from this perspective, the corpus-based and hand-crafted approaches to category assignment are complementary, not conflicting, where each can serve to validate and enrich the other.

6 CONCLUSION

The results of our experiments with Wikipedia are promisingly suggestive about the possibility of contextualizing ontological category structures via corpus-derived statistics. For example, the Wikipedia corpus reveals that the most common verb for the subject noun ”vampire” is ”hunt” (where the phrase ”vampires who hunt” occurs 4 times), indicating that in this pop-culture/fantasy-oriented context, a vampire is to be seen predominantly as a member of the category *hunt-agent*, or hunter. While one is unlikely to find such a categorization in an ontology like WordNet, or even Cyc, this is the most appropriate categorization in this context. Nonetheless, these results are hardly conclusive, for although large, Wikipedia is simply not large enough to provide the diversity of evidence needed to reliably derive a heterogeneous category membership. If a resource like Wikipedia lacks the necessary scale, surely this speaks to the futility of defining a context via a corpus?

We believe the answer to this dilemma lies not in ever-larger corpora (which may be too large to preserve the distinctive biases of a given context), but in the combination of different perspectives offered by the same corpus. We have described two different perspectives in this paper: the perspective of behaviour (captured via verb collocations) described in section 3, and the perspective of clustering (captured via group-word collocations) described in section 4. For instance, we know that Robot is the most representative member of the category *army-agent* in Wikipedia (with 63 examples), while army is itself a highly representative member of the category *attack-agent*. This suggests that Robot should also be a strong member of the category *attack-agent*. While Wikipedia records no uses of the collocation ”robot who|which|that attacks”, this joint perspective is sufficient evidence to support going to the web for this collocation. That is, the intuition that Robot is an *attack-agent* is consistent with the corpus, and thus the context, so the precise membership score can be determined using the larger context of the web.

Bootstrapping techniques like this should allow us to grow more heterogeneous category structures while respecting the ontological biases of the specific context. Once the deficiencies of relatively small corpora are addressed via such techniques, we expect to be better poised to fully explore the ramifications and opportunities of corpus-trained contextual ontologies.

ACKNOWLEDGEMENTS

We would like to thank Enterprise Ireland for supporting this research through a grant from the Commercialization Fund.

REFERENCES

[1] Dong, Z. and Dong, Q., HowNet and the Computation of Meaning, World Scientific, Singapore, 2006.

[2] Glucksberg, S. and Keysar,B., How Metaphors Work, *Metaphor and Thought (2nd edition)*, A. Ortony (Ed.), Cambridge University Press, 1993.

[3] Hofstadter,D., Fluid Concepts and Creative Analogies, Basic Books, 1995.

[4] Lakoff, G., *Women, Fire and Dangerous Things*, Chicago University Press, 1987.

[5] Lenat, D. and Guha, R. V., Building Large Knowledge-Based Systems, Addison Wesley, 1990.

[6] Miller, G. A., WordNet: A Lexical Database for English, *Communications of the ACM, Vol. 38 No.11*, 1995.

[7] Searle, J., Metaphor, *Metaphor and Thought (2nd edition)*, A. Ortony (Ed.), Cambridge University Press, 1993.

[8] Veale, T., Analogy Generation in HowNet, *The proceedings of IJCAI'2005*, the International Joint Conference on Artificial Intelligence, 2005.

[9] Wikipedia open-source encyclopaedia: *www.wikipedia.org.*

[10] F van Harmelen, L Serafini and H Stuckenschmidt, C-OWL: Contextualizing ontologies, *Proceedings of the 2nd International Semantic Web Conference*, 2003.

[11] Harris, Z., *Mathematical Structures of Language*, Wiley. 1968.

[12] Hindle, D., Noun classification from predicate-argument structures, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 268-275, 1990.

[13] P. Cimiano, A. Hotho, S. Staab., Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. *Journal of AI Research*, Volume 24: 305-339, 2005.

[14] Hearst, M., Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, pp. 539-545, 1992.