Creative Discovery in Lexical Ontologies

Tony VEALE
Dept. of Computer Science
Belfield,
Dublin, Ireland, D4
Tony.Veale@UCD.ie

Nuno SECO, Jer HAYES
Dept. of Computer Science
Belfield,
Dublin, Ireland, D4
{nuno.seco, jer.hayes}@UCD.ie

Abstract

Compound terms play a surprisingly key role in the organization of lexical ontologies. However, their inclusion forces one to address the issues of completeness and consistency that naturally arise from this organizational role. In this paper we show how creative exploration in the space of literal compounds can reveal not only additional compound terms to systematically balance an ontology, but can also discover new and potentially innovative concepts in their own right.

1 Introduction

Broad-coverage lexical knowledge-bases like WordNet (Miller *et al.*, 1990) generally contain a large number of compound terms, many of which are literal in composition. These compounds are undoubtedly included for a reason, yet the idea that literal compounds might actually be essential to WordNet's usefulness may strike some as heretical on at least two fronts: first, the lexicon is a finite resource, while the space of compounds is potentially infinite; and at any rate, literal compounds can be created as needed from purely compositional principles (Hanks, 2004). However, these retorts are valid only if we view WordNet as a dictionary, but of course it is much more than this. WordNet is a lexical ontology, and ultimately, ontologies derive a large part of their functionality from their structure.

So, while the meaning of literal compounds like *Greek-deity* and *animal-product* may well be predictable from compositional principles alone, such concepts still serve an important organizational role in WordNet by adding much needed structure to the middle ontology. Having conceded the importance of such compounds, one is forced to address the issues of completeness and consistency that then arise from their inclusion. Completeness suggests that we strive to include as many literal compounds as are sensible, if they enhance the organization of the ontology or if there is evidence that they are in common usage in the language. Systematicity is a related issue that arises when a group of existing compounds suggests that another should also exist for the ontology to be consistent. For instance, the existence of *Greek-deity*, *Greek-alphabet* and *Hebrew-alphabet* leads to the hypothesis that *Hebrew-deity* should also exist if WordNet is to be both consistent and symmetric in its treatment of different cultural groupings.

Indeed, because literal compounds like these arise from the yoking together of two different ontological branches into one, compounding represents an important contextualization device in the design of ontologies, allowing lexical elements to be logically grouped into clusters or families that share important dimensions of meaning. This clustering facilitates both automated reasoning by machines (such as the determination of semantic similarity based on taxonomic distance) and effective browsing by humans. Sometimes this yoking results in a compound that, following Boden (1990) and Wiggins (2003), deserves to be called "creative", because it exhibits both novelty and value. Novelty can be measured along either a psychological or a historical dimension, while utility is a reflection of the uses to which a compound can be put. For instance, a new compound may have utility as a clustering node when added to the middle ontology if its appropriate hyponyms can be identified. Alternately, a new compound may represent an alternate nominalization of an existing concept (e.g., see Vendler's (1967) insights about nominalization, and Lynott and Keane's (2003) application of these insights to compound generation).

In this paper we present a process of ontological exploration to identify those areas of the lexicon that can contribute to, and may in turn benefit from, the invention of new compound terms. Since the discovery of new compound terms is essentially a process of creative exploration, we frame our discussion within the theoretical framework of creative computation. Within this framework two approaches to validating new compounds are presented: internal validation determines whether the ontology itself provides evidence for the sensibility of a new compound,

while external validation uses web-search to find evidence that the compound already exists outside the ontology. We then go on to show how these different strategies create a validation gap that can be exploited to identify the small number of truly creative compounds that arise.

2 Exploring the Space of LMH Concepts

Creative discovery requires that we give structure to the space of possible concepts that we plan to explore. This is made somewhat easier if we consider the meaning of conceptual structures to be grounded in a semiotic system of meaning-creating oppositions. Given a starting structure, knowledge of allowable oppositions can then be used to transform this starting point into a variety of different conceivable structures, some of which may be novel and possess value on a particular utility scale.

The notion of opposition employed here is much broader than that of antonymy. For our purposes, contextual oppositions exist between terms that compete to fill a given dimension of the same concept. For instance, *Greek*[1] and *Hindu* can each be used to differentiate the concept *deity* along a *culture* dimension, and so, in the context of *deity*, both are opposed. However, this is a contextual opposition that, unlike the role of antonymy, does not constitute part of the meaning of either concept. WordNet is a rich source of explicit antonymous oppositions, but contextual oppositions must be inferred from the structure of the ontology itself and from existing compounds.

Fortunately, WordNet contains many instances of literal modifier-head terms, such as "pastry crust" and "Greek alphabet". The concepts denoted by these compound terms, or LMH concepts for short, have the lexical form M-H (such as *pizza-pie* or *prairie-dog*) and express their literality in two ways. First, they must be stored in the WordNet ontology under an existing sense of the lexeme H; for instance, *pizza-pie* is actually stored under the hypernym *pie*. Secondly, the gloss for the concept M-H should actually contain the lexeme M or some synonym of it. Thus, while *Greek-alphabet* is a LMH (it literally is a kind of *alphabet*, and it is literally *Greek*), neither *monkey-bread* (which is not literally a kind of *bread*) nor *Dutch-courage* (which is not literally *Dutch*) is a LMH concept.

2.1 A Framework for Creativity

We use the terminology of Wiggins (2003) to frame our discussion of creative exploration. Wiggins, following earlier work by Boden (1990), formalizes the creative exploration process using the following abstractions:

C - the realm of concepts that is being explored

R - the set of rules for forming concepts and conversely, deconstructing existing ones

T - the transformational rules that generate new concepts via R

E - the evaluation mechanism that ascribes value or utility to these new concepts

In applying these terms to creativity in WordNet, we introduce the following refinements:

C_W - the subset of C described explicitly in WordNet as synsets

C* - the set of LMH concepts in C_W considered as a starting point for creative exploration

R* - the subset of R needed to construct and deconstruct LMH compounds in C*

T* - the subset of T needed to hypothesize new LMH concepts for R* to construct

So for our current purposes, we define C* as the set of LMH concepts in WordNet, and R* as the compositional criteria used to identify and decompose existing LMH entries and to construct new ones by concatenating an appropriate M and H term pair. However, to define T*, we first need to consider how taxonomic differentiation is used to create LMH concepts in the first place.

3 Domain Differentiation

LMH concepts exist in WordNet to differentiate more general concepts in meaningful taxonomic ways. For instance, the LMH concepts *Greek-alphabet*, *Hebrew-alphabet* and *Roman-alphabet* each serve to differentiate the concept *alphabet*. This is a useful ontological distinction that contributes to the definition of individual letter concepts like *Alpha, Beta* and *Gimel*. Since we can represent this specialization pattern via a differentiation set $D_{alphabet}$ as follows:

$$D_{alphabet} = \{Greek, Hebrew, Roman\}$$

More generally, the differentiation set of a concept H comprises the set of all concepts M such that the LMH concept M-H is in C*. Thus we have:

$$D_{deity} = \{Hindu, Roman, Greek, \ldots\}$$

$$D_{architecture} = \{Greek, Roman, \ldots\}$$

$$D_{calendar} = \{Muslim, Jewish, Hebrew, \ldots\}$$

[1] To avoid later confusion with set notion, we denote WordNet senses not as synsets but as italicized terms .

We use D to denote the set of all differentiation sets that are implied by C*, allowing us to define the absolute affinity between two modifier terms c_1 and c_2 in terms of differentiation as follows:

$$A_{abs}(c_1, c_2) = |\{x \in D: c_1 \in x \wedge c_2 \in x\}| \quad (1)$$

This simply counts the number of base concepts that c_1 and c_2 can both differentiate. We thus define the relative affinity between two modifier terms c_1 and c_2 as follows:

$$A_{rel}(c_1, c_2) = |\{x \in D : c_1 \in x \wedge c_2 \in x\}| \quad (2)$$
$$/ \ |\{x \in D : c_1 \in x \vee c_2 \in x\}|$$

A relative affinity of 1.0 means that both terms differentiate exactly the same concepts in WordNet. It follows that the higher the relative affinity between c_1 and c_2, then the greater the likelihood that a concept differentiated by c_1 can also be differentiated by c_2, while the higher the absolute affinity, the more reliable this likelihood estimate becomes. Affinity thus provides an effective basis for formulating the transformation rules in T*.

We should naturally expect near-synonymous modifiers to have a strong affinity for each other. For instance, *Jewish* and *Hebrew* are near-synonyms because WordNet compounds *Jewish-Calendar* and *Hebrew-Calendar* are themselves synonymous. This is clearly a form of contextual synonymy, since *Jewish* and *Hebrew* do not mean the same thing. Nonetheless, their affinity can be used to generate new compounds that add value to WordNet as synonyms of existing terms, such as *Jewish-alphabet, Hebrew-Religion*, and so on.

Recall that literal compounds represent a yoking together of two or more ontological branches. In exploring the space of novel compounds, it will be important to recognize which branches most naturally form the strongest bonds. Another variant of affinity can be formulated for this purpose:

$$A_{domain}(x, y) = |D_X \cap D_X| \quad (3)$$

For instance, A_{domain}(*sauce, pizza*) = 2, since in WordNet the modifier overlap between the *pizza* and *sauce* domains is {*anchovy, cheese*}.

4 Creative Exploration in the LMH Space

We consider as an exploratory starting point any LMH concept M-H in C*. We can transform this into another concept M'-H by replacing M with any M' for which:

$$M' \in \{x \mid x \in D - \{D_H\} \wedge M \in x\} \quad (4)$$

This formulation may suggest a large range of values of M'. However, these candidates can be sorted by A_{rel}(M, M'), which estimates the probability that a given M'-H will later be validated as useful. One rule in T* can now be formulated for our further consideration:

$$T^*: \ M_1\text{-}H_1 \wedge M_1\text{-}H_2 \wedge M_2\text{-}H_1 \Rightarrow M_2\text{-}H_2 \quad (5)$$

This rule allows the LMH space to be explored via a process of modifier modulation. Suppose we choose *Greek-deity* as a starting point. Since M = *Greek* and H = *Deity*, we can choose M' from any set other than D_{deity} that contains *Greek:*

$D_{alphabet}$ = {*Hebrew, Greek, Roman*}

D_{deity} = {*Greek, Roman, Norse, Hindu, ...*}

These differentiation patterns suggest that new compounds can meaningfully be created by yoking the ontological branches of *alphabet* and *deity* together. Thus, from $D_{alphabet}$ we can choose M' to be either *Hebrew* or *Roman*, leading to the creation of the LMH concepts *Hebrew-deity* and *Roman-deity*. One of these, *Roman-deity*, already exists in C*, but another, *Hebrew-deity* is novel in a way that Boden terms psychologically or P-Creative, inasmuch as it is neither in C_W nor C*. It may thus be of some value as a hypernym for existing WordNet concepts like *Yahwe* and *Jehovah*.

Rule (5) is a general principle for ontological exploration in the space of compound terms. Consider the compound *software-engineering,* which, following (5), is suggested by the joint existence in WordNet of the concepts *software-engineer, automotive-engineer* and *automotive-engineering*. While this particular addition could be predicted from the application of simple morphology rules, the point here is that a single exploration principle like (5) can obviate the need for a patchwork of such simple rules.

Of course, one can imagine rules other than (5) to exploit the regularities inherent in WordNet definitions. For instance, consider the sense *gasoline-bomb*, which WordNet glosses as: "a crude incendiary bomb made of a bottle filled with flammable liquid and fitted with a rag wick". By determining which definite description in the gloss conforms to the modifier – in this case it is "flammable liquid" – other modifiers can be found that also match this description. Thus, the new concepts *methanol-bomb* and *butanol-bomb* can be generated, and from this the creative concept *alcohol-bomb* can be generalized. However, each

strategy raises its own unique issues, so for now we consider a T* comprising (5) only.

4.1 The Evaluation Mechanism E

For purposes of ascribing value or ontological utility to a new LMH concept M'-H, the concept must first be placed into one of the following categories:

a) M'-H already exists in C* is thus ascribed zero value as an addition to C*.

b) M'-H does not exist in C* but does exist in C_W, and thus corresponds to an existing non-literal concept (such as *monkey-bread*). While it may have value if given a purely literal reading, it cannot be added to C_W without creating ambiguity, and so has zero value.

c) Using R*, M'-H can be seen to describe a non-empty class of existing concepts in C_W, and would thus have value as either a synonym (when this set is a singleton) or as a new organizing super-type (when this set is a severalton). In this case, we say that M'-H has been *internally validated* against C_W.

d) Using a textual analysis of a large corpus such as the World-Wide-Web, M'-H is recognized to have a conventional meaning in C even if it is not described in C_W. In this case, we say that M'-H has been *externally validated* for inclusion in C_W. The fact that M'-H is novel to the system but not to the historical context of the web suggests that it is merely a psychologically or P-Creative invention in the sense of Boden (1990).

e) M'-H is recognized to have a hypothetical or metaphoric value within a comprehension framework such as conceptual blending theory (e.g., see Veale *et al*. 2000), mental space theory, etc. In this case, M'-H may truly be a historically or H-Creative invention in the sense of Boden (1990).

In general, a new compound has value if its existence is suggested by, but not recognized by, the lexical ontology. As noted in the introduction, this value can be realized in a variety of ways, e.g., by automatically suggesting new knowledge-base additions to the lexical ontologist, or by providing potentially creative expansions for a user query in an information retrieval system (see Veale, 2004).

4.2 Validating New Concepts

The evaluation strategies (c) and (d) above suggest two ways of validating the results of new compound creation: a WordNet-internal approach that uses the structure of the ontology itself to provide evidence for a compound's utility, and a WordNet-external approach that instead looks to an unstructured archive like the web. In both cases, a new compound is validated by assembling a support set of precedent terms that argue for its meaningfulness.

4.2.1 Internal Validation

The internal support-set for a new compound M-H is the set of all WordNet words *w* that have: (i) at least one sense that is a hyponym of a sense of H; and (ii) a sense that contains M or some variant of it in its gloss. For instance, the novel compound "rain god" is internally validated by the word set {"Thor", "Parjanya", "Rain giver"}.

When *w* is polysemous, two distinct senses may be used, reflecting the fact that M-H may be metonymic in construction. For instance, the compound "raisin-wine" can be validated internally by the polysemous word "muscatel", since one sense of "muscatel" is a kind of wine, and another, a kind of grape, has a WordNet gloss containing the word "raisin". From this perspective, a "raisin wine" can be a wine made from the same grapes that raisins are made from. Likewise, the compound "Jewish robot" can be validated by simultaneously employing both senses of "Golem" in WordNet, which defines "Golem" as either a Jewish mythical being or as a robotic automaton.

Creative products arise when conceptual ingredients from different domains are effectively *blended* (see Veale and O'Donoghue, 2000). It follows that a creative product can be validated by deblending it into its constituent parts and determining whether there is a precedent for combining elements of these types, if not these specific elements. We can thus exploit this notion of deblending to provide internal validation for new compounds. For instance the WordNet gloss for pizza lists "tomato sauce" as an ingredient. This suggests we can meaningfully understand a compound of the form M-*pizza* if there exists a compound M-*sauce* that can be viewed as a replacement for this ingredient. Generalizing from this, we can consider a new compound M_1-H_1 to be internally validated if H has a sense whose gloss contains the compound M_2-H_2, and if the ontology additionally contains the concept M_1-H_2. It follows then that the novel compounds *apple-pizza, chocolate-pizza, taco-pizza,* and *curry-pizza* will all be internally validated as meaningful (if not necessarily enjoyable) varieties of pizza.

4.2.2 External Validation

In contrast, the external validation set for a compound M-H is the set of distinct documents that contain the compound term "M H", as acquired using a web search engine. For instance, given the WordNet concepts *naval-engineer*, *software-engineer* and *naval-academy*, rule (5) generates the hypothesis *software-academy,* which cannot be validated internally yet which retrieves over 1000 web documents to atest to its validity.

This web strategy is motivated by Keller and Lapata's (2003) finding that the number of documents containing a novel compound reliably predicts the human plausibility scores for the compound.

Nonetheless, external validation in this way is by no means a robust process. Since web documents are not sense tagged, one cannot be sure that a compound occurs with the sense that it is hypothesized to have. Indeed, it may not even occur as a compound at all, but as a coincidental juxtaposition of terms from different phrases or sentences. Finally, even if found with the correct syntactic and semantic form, one cannot be sure that the usage is not that of a non-native, second language learner.

These possibilities can be diminished by seeking a large enough sample set, but this has the effect of setting the evidential bar too high for truly creative compounds. However, another solution lies in the way that the results of external validation are actually used, as we shall later see.

4.2.3 Validating New Synonyms

Many of the compounds that are validated either by internal or external means will be synonyms of existing WordNet terms. As such, their creative value will not represent an innovative combination of ideas, but rather a creative use of paraphrasing. The nature of (5) makes it straightforward to determine which is the case.

In general, when M_1-H_1 and M_2-H_1 are themselves synonyms, then M_2-H_2 will be a synonym of M_1-H_2. For instance, from the combination of *applied-science, engineering-science* and *applied-mathematics*, we can generate from (5) the new compound *engineering-mathematics*. This compound cannot be validated internally, but since it retrieves more than 300,000 documents from the web, this is enough to adequately atest to its meaningfulness. Now, since *applied-science* and *engineering-science* are synonymous in WordNet, we can conclude that *engineering-mathematics* and *applied-mathematics* are themselves synonymous also.

4.3 Creativity in the Validation Gap

The difference between internal and external validation strategies can be illuminating. Internal validation verifies a compound on the basis that it *could meaningfully* exist, while external validation verifies it on the basis that it *does actually* exist in a large corpus. Therefore, if a compound can be validated externally but not internally, it suggests that the concept may by P-Creative. In contrast, if the compound can be validated internally but not externally, it suggests that the compound may be H-Creative and represent a genuine historical innovation (if only a lexical one, and of minor proportions).

For instance, the new compound "sea dance" (analogous to "rain dance") cannot be validated internally, yet can be found in over 700 internet documents. It thus denotes a P-Creative concept. In contrast, the compound "cranial vein" yields no documents from a web query (on AltaVista), yet can be internally validated by WordNet via the word-concept *Diploic-Vein*, a blood vessel that serves the soft tissue of the cranial bones. Likewise, the compounds "chocolate pizza", "taco pizza" and many more from the yoking of D_{pizza} and D_{sauce} can all be validated externally via hundreds of different web occurrences, and so represent P-Creative varieties of pizza. However, compounds like "Newburg pizza" (a pizza made with lobster sauce) and "wine pizza" (a pizza made with wine sauce) can only be validated internally and are thus candidates for H-Creative innovation.

5 Large-Scale Evaluation

A large scale evaluation of these ideas was conducted by exhaustively applying the T* rule of (5) to the noun taxonomy of WordNet 1.7. To better see the effect of affinity between modifiers, Table 1 ranks the results according to the measure A_{abs} from (1).

A_{abs}	1	2	3
No. compounds generated	941,841	22,727	2,175
% H-Creative	0.49%	0.63%	1.38%
% P-Creative	35.65%	33.77%	34.57%
% Conflations	0.10%	0.10%	0.05%
%Indeterminate	63.76%	65.49%	64.00%

Table 1: Number of compounds created, and their assessment, for each affinity level.

Conflations are terms that exist both as compounds

and as conflated lexical atoms. For instance, while the compound "bull dog" may not exist in WordNet, its conflation "bulldog" does. Compound discovery is thus a useful means of re-expanding these conflations when it is meaningful to do so.

As one might expect, lower affinity levels allow greater numbers of new compounds to be created. Interestingly, however, Table 1 suggests that as the affinity threshold is raised and the number of compounds lowered, the creativity of these compounds increases, as measured by the relative proportion of H-Creative terms that are generated.

Generating compound terms in a lexical ontology is a creative process that demands rigorous validation if the ontology is not to be corrupted. Of the two strategies discussed here, external validation is undoubtedly the weaker of the two, as one should be loathe to add new compounds to WordNet on the basis of web evidence alone. However, external validation does serve to illustrate the soundness of internal validation, since 99.51% of internally validated concepts (at $A_{abs} = 1$) are shown to exist on the web. It follows then that the *absence* of external validation yields a very conservative basis for assessing H-Creativity. Web validation is perhaps better used therefore as a means of rejecting creative products than as a means of discovering them. In fact, when used as a reverse barometer in this way, the inevitable errors that arise from web-based validation serve only to make the creative process more selective.

6 Conclusions and Future Work

We are currently considering ways of broadening the scope of internal validation while maintaining its conceptual rigour. This should counter-balance the high rejection rate caused by an overly conservative external validation process, and thereby allow us to identify a higher percentage of H-creative products. As shown with the "pizza" examples of section 4.3, we have already begun to explore the possibilities of validation latent in the WordNet ontology itself. So while the use of web content for external validation suggests that creative discovery has a role to play in producing and expanding web queries, internal validation remains our central thrust, leading to what we hope will be a new, more creative model of the thesaurus.

In grounding our discussion in the creative framework of Boden (1990) and its formalization by Wiggins (2003), we have placed particular emphasis on the labels P-Creative and H-Creative. However, the empirical results of section 5 suggest that this binary categorization may be overly reductive, and that a more gradated system of labels is needed. For instance, the novel compounds *computer-consultant* and *handwriting-consultant* are both created from a yoking of the domains *expert* and *consultant*, and because each is externally validated, each is considered P-Creative.

However, while only a handful of documents can be marshalled to support *handwriting-consultant,* the amount of web evidence available to support *computer-consultant* is vast. So it is wrongheaded to consider both as equally P-Creative and lacking in H-Creativity, since the dearth of existing uses suggests *handwriting-consultant* has far greater novelty. Perhaps what is needed then is not a binary categorization but a continuous one, a numeric scale with P- and H-Creativity as its poles. This scale would function much like the continuum used by (MacCormac, 1985) to separate banal metaphors (which he dubbed *epiphors*) from creative ones (or *diaphors*).

References

M. A. Boden. 1990. ***The Creative Mind: Myths and Mechanisms*** New York: Basic Books.

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

F. Keller, and M. Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. ***Computational Linguistics***

Lynott, Dermot and Mark Keane, 2003. The role of knowledge support in creating noun-noun compounds. In ***the proceedings of the 25th Conference of the Cognitive Science Society***

E. R. MacCormac. 1985. A Cognitive Theory of Metaphor. Cambridge, MA: MIT Press.

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

G. Wiggins. 2003. Categorizing Creative Systems. ***In the proceedings of the 3rd Workshop on Creative Systems, IJCAI'03*** Acapulco, Mexico.

T. Veale and D. O'Donoghue. 2000. Computation and Blending. ***Cognitive Linguistics***; 11(3/4): 253— 281.

T. Veale. 2004. Creative Information Retrieval. ***In the proceedings of CICLing 2004*** A. Gelbukh, ed. LNCS 2945, Springer: Berlin.

Z. Vendler. 1967. ***Linguistics and Philosophy*** Ithaca, New York: Cornell University Press.