Concept Creation in Lexical Ontologies

Nuno Seco, Tony Veale, Jer Hayes

University College Dublin
Department of Computer Science
Dublin 4, Ireland
{nuno.seco, tony.veale, jer.hayes}@ucd.ie

Abstract

The compositional mechanisms involved in the comprehension and creation of concepts is of much interest to the communities studying Cognitive Science and Artificial Intelligence. Nevertheless, comprehension has been largely studied while the creation or production of novel concepts has been somewhat forgotten. We present a model for concept generation using a well known lexical ontology — WordNet — along with the results of our experiments that evaluate the creative characteristics of the generated concepts. We also explain how these ideas may be applied to other areas of research, namely to Information Retrieval systems.

1. Introduction

Large-scale ontologies are necessarily incomplete, whether due to designer oversight or the dynamic nature of the domain being ontologized. For example, WordNet (Miller et al., 1990) is an ontology that represents a synchronic snapshot of the English lexicon, an inherently diachronic system. As such, large ontologies can only hope to capture a salient selection of the concepts that should be represented. For WordNet, the selection criterion is determined by conventional word usage, but many concepts that can profitably be represented are omitted. These omissions lead to holes and asymmetries in the ontology that can significantly mislead automated reasoning systems that are sensitive to the organization of the ontology, e.g., in performing text categorization or in calculating inter-word similarity measures.

Fortunately, the concepts that are explicitly ontologized can serve as a guide to many of the concepts that are omitted. In this paper, we present a model of exploratory creativity that uses the existing WordNet ontology as a basis for inducing the concepts that WordNet appears to lack and which should profitably be added. In the first phase of discovery, the compound concepts of the existing lexical ontology are analysed and deconstructed, to yield a vocabulary of atomic elements from which new concepts can be constructed. In itself, this vocabulary defines an excessively large space to explore without explicit signposts. So in the second phase of discovery, a simple grammar is also extracted from the existing ontology, to identify a sweet-spot in the space of possible concepts that can feasibly be explored. In the third phase, new compound concepts are generated by applying the grammar to the vocabulary. In the fourth and final phase, each new concept is validated to ensure ontological utility, whereby those of value are added to the ontology and those without value are rejected.

Naturally, the value of an ontology like WordNet is significantly compromised if nonsense concepts are introduced. The validation phase is thus the most crucial of the entire discovery process and the one deserving of the most computational resources. In this paper we describe two complementary validation procedures. The internal validation process attempts to situate a novel concept within the ontology using existing concepts as a guide, for if read appropriately, the existing ontology can be highly suggestive of the holes that need to be filled. In contrast, the external validation process uses the World Wide Web as a repository of previously lexicalized concepts that can be queried using an Internet search engine[1]. Internal validation measures the contribution a new concept makes to the organization of the ontology by increasing the accessibility of existing concepts, while external validation measures the necessity of a new entry to the ontology based on its use in the wider language community of the WWW. As one might expect, many concepts that cannot be validated internally can be validated by recourse to the vastness of the WWW. In creativity terms as defined by Margaret Boden (Boden, 1990), such concepts are merely P-creative, demonstrating psychological novelty only in the narrow sense of being new to the ontology itself. More interesting, however, is the realization that some concepts that can be validated internally cannot be validated externally. These concepts deserve to be dubbed H-creative, demonstrating a historical originality that suggests ontologies like WordNet can sensibly be used as the basis of creative linguistic systems.

This paper will report a detailed empirical analysis of our experiments with concept creation in WordNet[2], with pointers to the applicability of these ideas to text-based applications like information retrieval.

2. Concept Creation

The creation and interpretation of compounds has been a focal topic of study in cognitive science and artificial intelligence. Several decades of research have been devoted to the study of nominal compounds, however Lynott and Keane (Lynott and Keane, 2003) point out that much of this research has addressed the comprehension of compounds leaving the creation somewhat overlooked. Despite this trend, some research regarding compound generation has been conducted (see (Lynott and Keane, 2003; Pereira, 2003)).

Lynott and Keane put forward a model for compound

[1] In the experiments reported in this paper we use the AltaVista search engine — http://www.altavista.com

[2] In the experiments reported in this paper we use WordNet 1.6

production based on object descriptions, a model that mimics the process of nominalization described by Vendler (Vendler, 1967). According to them this process can be automated by *"finding the minimal subset of terms whose meaning will accurately and unambiguously convey the given meaning.*". So for example, if the given descriptions are:

1. A wine that is made from grapes and contains alcohol.
2. A wine that is made from apricots and contains alcohol.

then for (1) the unique word *wine* will suffice to convey the intended meaning, on the other hand for (2), an extra word is needed yielding *apricot wine*. In the above computations several types of knowledge are used that allow us to consider the above subsets as reasonable indexes of the overall description. Lynott identifies three knowledge types that are likely to be relevant:

- pragmatic knowledge.
- world knowledge.
- syntactic knowledge.

The experiments conducted by his research examined the effects of world and syntactic knowledge in the process of compound production by human participants.

Our research on concept generation differs from that of Lynott in the sense that we are not interested in investigating what types of knowledge are involved in the process of concept production, but on reusing the implicit knowledge already contained in existing concepts and in WordNet as a whole. Hence known compounds (contained in WordNet) are used to create novel and meaningful compound concepts. The plausibility of each compound is then evaluated internally through the use of the WordNet taxonomy, or externally through the use of a web search engine.

3. A Model for Concept Creation

The system first starts out by obtaining all clusters of existing compounds from WordNet. In this context a cluster of existing compounds is a group of compounds that are hyponyms of the same concept and that share the same head. Formally, such a cluster may be defined as:

$$\begin{aligned}\mathcal{C} = \{&M_1H, .., M_nH \mid \forall i,j \in \{1..n\} \\ &M_iH\ ISA\ Hypernym \wedge \\ &M_jH\ ISA\ Hypernym \wedge \\ &M_i \neq M_j\ if\ i \neq j\}\end{aligned} \quad (1)$$

In definition 1, *M* and *H* denote the modifier and head of a compound, respectively. So if the compound is *"monkey bread"* then according to the above definition we have that M_i = *monkey*, H = *bread* and *Hypernym* = *edible fruit*. An example of such a cluster is <<*snake god, sun god, earth god, war god, sea god*>>[3] which are all hyponyms of *deity* and all share the common head *god*.

[3]We use the double angle bracket notation to avoid confusion with the synset representation of WordNet which employs curly brackets.

After obtaining the clusters the system detects intersections between the modifiers of the compounds of each cluster. Consider the cluster <<*rain dance, sun dance, ghost dance, war dance, snake dance*>> in which the elements are all hyponyms of *ritual dancing* and the *deity* cluster presented in the preceding paragraph. As can be observed, some elements of both clusters share common modifiers, such as *sun*, *war* and *snake*. These elements constitute the Modifier Intersection Set ($\mathcal{MIS}$) for the given clusters. Formally we have;

$$\mathcal{MIS} = \mathcal{MS}_1 \cap \mathcal{MS}_2 \quad (2)$$

where $\mathcal{MS}_1$ and $\mathcal{MS}_2$ represent the Modifier Set (MS) of each cluster $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively.

After discovering the $\mathcal{MIS}$ for all pairs of clusters the system may speculate about the existence of new compound concepts. Returning to the *deity* and *ritual dancing* clusters, the system would suggest the creation of *earth dance* and *sea dance* as hyponyms of *ritual dancing*; and *rain god* and *ghost god* as hyponyms of *deity*. Thus, we say that if $\mathcal{C}_1$ is a cluster of compounds that share the head $\mathcal{H}_1$ with a modifier set $\mathcal{MS}_1$ and $\mathcal{C}_2$ is a cluster of compounds that share the head $\mathcal{H}_2$ with a modifier set $\mathcal{MS}_2$ and if the $\mathcal{MIS}$ between $\mathcal{C}_1$ and $\mathcal{C}_2$ is non-empty then the set $\mathcal{MS}_1 - \mathcal{MIS}$ may be used to differentiate $\mathcal{H}_2$ and the set $\mathcal{MS}_2 - \mathcal{MIS}$ to differentiate $\mathcal{H}_1$. Hence, speculation is constrained to the extracted grammar allowing efficient exploration of the search space.

4. Validation Phase

Subsequently to the actual creation process is the validation phase, where the plausibility of the concept is determined. Here we try to determine if the newly generated concept is reasonable and sensible either by using the taxonomic structure of WordNet (internal validation) or from information available on the internet (external validation) through AltaVista.

Internal validation requires evidence from the taxonomic structure that the head of the created compound may be sensibly modified with a particular word. The evidence is determined by looking at all hyponyms of the head and checking if their glosses explicitly refer to the modifier, a technique much in the same vein as the one used in (Veale, 2003). If such a concept is found then it is reasonable to accept the new concept, otherwise it is rejected. Revisiting our example, consider the concept *rain god*. This concept is internally validated by WordNet because we can find at least one hyponym concept of *god* (the head of the compound) that mentions *rain*. In this particular case we actually find 3 concepts:

1. rain giver — an epithet for Jupiter.
2. thor — (Norse mythology) god of thunder and rain and farming.
3. parjanya — god of rain.

Thus, by observing the definitions of these concepts it seems reasonable to consider *rain god* as meaningful. It should be noted that *rain god* can be introduced into the

taxonomy as a hypernym for the concepts that hold the required evidence (*rain giver*, *thor*, *parjanya*).

Unfortunately, the WordNet glosses do not always hold the required evidence. In order to overcome this difficulty the internet is also used. Previous research, conducted by Keller and Lapata (Keller and Lapata, 2003), has shown that the number of results obtained for a compound query (e.g. noun-noun compound) reliably predict human plausibility judgments. We assume that a compound concept is validated if a query representing that concept when sent to AltaVista returns more then 10 hits from distinct sites. (Note that queries are for exact matches, which means that the compound concept must be enclosed between quotation marks.) The assumption here is that if a bigram can be found at least ten times in different and unrelated documents then it can be considered meaningful. Take the generated concept *sea dance*; at the time of writing of this paper AltaVista finds 721 documents containing the specified bigram which strongly suggests its significance. Nevertheless, using the internal validation scheme presented above, its validity is not determined.

In the realm of creativity, and considering Margaret Bodens' view, we argue that concepts that are validated internally and that are not found on the web correspond to H-creative concepts. An example of such a concept is *cranial vein*; the result of an internet query does not assure the plausibility (in terms of the number of results) of the new concept, however this term is validated internally when we consider the concept *diploic vein* which is defined as *"one of the veins serving the spongy part of the cranial bones"*. If one were to analyze this gloss in the light of the *minimal subset* definition of Lynott and Keane, it can be argued that *cranial vein* is indeed a good representative of the intended meaning. When candidate concepts are validated using the web we must consider the concepts as P-creative, independently of the internal validation result.

5. Empirical Studies

In order to evaluate the creative potential of our system several experiments were conducted. The focus of these initial experiments is to understand how the size of $\mathcal{MIS}$ between two clusters influences the quality of the generated concepts. Therefore we separate the new concepts into eight groups, the group in which a concept is placed depends on the size of the $\mathcal{MIS}$ that generated that concept. Revisiting the example of the previous section, the size of the $\mathcal{MIS}$ between the *deity* cluster and the *ritual dancing* cluster is 3, hence all concepts generated due to these clusters will belong to group 3. After allocating all concepts to their respective group we evaluate there creativity based on the two validation schemes presented earlier. All generated concepts that were found to already exist in WordNet were immediately discarded and not considered in the validation process. It should also be noted that some concepts are generated more than once, thus one concept can have more than one $\mathcal{MIS}$ associated, in these cases we allocate it to the smallest $\mathcal{MIS}$. Our results are presented in table 1.

Table 1 in addition to showing the number of generated concepts and the percentage considered P-Creative or H-Creative, we also calculate the percentage of synonyms and the percentage of concepts that were not validated, neither by internal nor external validation. We consider synonyms to be concepts that already exist in WordNet but with a different morphological representation[4]. Wordnet usually represents its compound concepts by placing an underscore (_) between the modifier and head (e.g. *monkey_bread*), but this representation is not consistent throughout the ontology, so care must be taken in order not to mistake some of the generated concepts as novel. For example, our system generates the compound *bull_dog* but since the word *bulldog* already exists we consider *bull_dog* a synonym.

6. Conclusions and Future Work

Observing the data presented in table 1 we see that the size of the $\mathcal{MIS}$ influences the number of generated concepts. This seems sound, since the larger the $\mathcal{MIS}$ the more restrictive the grammar becomes, thus loosing generative potential. This initial study shows that the proposed algorithm for compound concept generation performs reasonably well, with an average of 42.49%[5] of the generated compounds being meaningful (either H-Creative, P-Creative or Synonymous). As expected, H-Creative concepts represent a very small portion of the generated concepts, but since the plausibility of these concepts can be explained by our system, and not by an Internet Query we find the results very appealing.

We feel that such a technique may be amenable to Information Retrieval systems that make use of query expansion. A user's query may be expanded with concepts, that otherwise wouldn't be considered (due to their rare occurrence) using statistical algorithms. This technique may be used to complement existing Query Expansion strategies boosting recall without diminishing precision.

Future work will consist of improving the validation techniques in order to find relevant concepts (that can be sanely justified by a computational knowledge driven inference) that are at the moment being classified as indetermined. Another aspect that will deserve our attention is the identification of other compounds, during external validation, that do not exist in WordNet but since they appear in a known context can be used to augment it in a profitable manner.

7. References

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

Keller, Frank and Maria 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 *Proceedings of the Twenty-Fifth Conference of the Cognitive Science Society*.

Miller, George, Richard Beckwith, Fellbaum Christiane, Derek Gross, and Katherine J. Miller, 1990. Introduction

[4] WordNet already employs such a presumption, consider the synset — {groundcover, ground_cover}

[5] Group 8 was not considered in the calculation.

Group	Concepts	H-Creative	P-Creative	Synonyms	Indetermined
1	941841	0.49%	35.65%	0.10%	63.77%
2	22727	0.63%	33.77%	0.10%	65.49%
3	2175	1.38%	34.57%	0.05%	64.00%
4	250	0.40%	66.40%	0.00%	33.20%
5	178	1.12%	24.72%	0.00%	74.16%
6	45	0.00%	55.56%	0.00%	44.44%
7	0	N/A	N/A	N/A	N/A
8	2	0.00%	100.00%	0.00%	0.00%

Table 1: Evaluation of Generated Concepts.

to wordnet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235 – 244.

Pereira, Francisco, 2003. Experiments with free concept generation in divago. In *Proceedings of the Third Workshop on Creative Systems*.

Veale, Tony, 2003. The analogical thesaurus: An emerging application at the juncture of lexical metaphor and information retrieval. In *proceedings of IAAI 2003, the 2003 International Conference on Innovative Applications of Artificial Intelligence*.

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