Analogical Reasoning with a Synergy of WordNet and HowNet

Hyesook Kim,	Shanshan Chen,	Tony Veale,
School of Computer Science,	School of Computer Science,	School of Computer Science,
University College Dublin,	University College Dublin,	University College Dublin,
Dublin D4, Ireland.	Dublin D4, Ireland.	Dublin D4, Ireland.
Hyesook.Kim@UCD.ie	Shanshan.Chen@UCD.ie	Tony.Veale@UCD.ie

Abstract

WordNet and HowNet are large-scale lexical resources that adopt complementary perspectives toward semantic representation. WordNet is *differential*, inasmuch as it provides a rich taxonomic structure but little in the way of explicit propositional content. HowNet is *constructive*, and dedicates its representational energies to the explicit definition of relational structures for each lexico-conceptual entry. For purposes of analogy, no one approach is best. Rather, a synergy of both is required, since analogy is a knowledge-hungry process that demands both taxonomic richness *and* causally descriptive propositional structure. In this paper we consider how such a synergy might be achieved, and how it can be exploited to support a robust model of analogical reasoning.

1 Introduction

Analogy is a knowledge-hungry process that exploits a conceptual system's ability to perform controlled generalization in one domain and re-specialization into another. The result is a taxonomic leap within an ontology that transfers semantic content from one term onto another. While all taxonomies allow vertical movement, a system must fully understand the effects of generalization on a given concept before any analogy or metaphor can be considered either deliberate or meaningful. For instance, an analogizer must be capable of generalizing over the causal structure of source and target concepts, since research has shown that the most satisfying analogies are those that operate at the causal level of representation. This importance of causality follows from analogy's usefulness in offering a deep explanation of a poorly understood phenomenon in terms of a well understood one (see Falkenhainer *et al.,* 1989; Veale and Keane, 1997). For instance, the analogy *Atom as miniature Solar-System* is satisfying because both source and target concepts are causally structured around the notion of rotation. It follows then that to support analogy, a taxonomy must provide a sound basis of abstracting over the causal structure and functional behaviour of concepts.

When comparing agents or artefacts, this causality can be abstracted out by considering the functional or behavioural commonality between target and source: a surgeon can be meaningfully described as a repairman since both occupations have the function of restoring an object to an earlier and better state; a footballer can be meaningfully described as a gladiator or a warrior since each exhibits competitive behaviour; and a scalpel can be compared to a sabre, a sword or a cleaver since each has a cutting behaviour; and so on.

Theories of metaphor and analogy are typically based either on structure-mapping (again see Falkenhainer *et al.,* 1989; Veale and Keane, 1997) or on abstraction (e.g., see Hutton, 1982; Fass, 1988; Way, 1991; Veale, 2003, 2004). While the former is most associated with analogy, the latter has been a near-constant in the computational treatment of metaphor. Structure-mapping assumes that the causal behaviour of a concept is expressed in an explicit, graph-theoretic form so that unifying sub-graph isomorphisms can be found between different concepts. In contrast, abstraction theories assume that analogous concepts, even when far removed in ontological terms, will nonetheless share a common hypernym that captures their causal similarity. Thus, we should expect an analogous pairing like *surgeon* and *butcher* to have different immediate hypernyms but to ultimately share an abstraction like *cutting-agent* (see Veale, 2003,2004).

However, the idea that a standard ontology will actually provide a hypernym like *cutting-agent* seems convenient almost to the point of incredulity. The problem is, of course, that as much as we want our ontologies to anticipate future analogies and metaphors with these pro-active categorizations, most ontologies simply do not possess terms as prescient as these. This is the question we address in this paper: if we assume that our ontologies lack these structures, can we nonetheless enable them to be added via automatic means? We argue that we can, by looking not to purely conceptual representations but to integrated natural language representations derived from multiple sources and multiple languages. By integrating WordNet (see Miller, 1995) and HowNet (see Dong, 1998; Gan and Wong, 2000; Carpuat *et al.*, 2002; Wong, 2004) – two lexical ontologies built around different approaches to semantic structure – we hope to create a representational synergy that is better suited to the demands of analogical reasoning.

2 Abstraction Theories of Analogy

That analogy and metaphor operate across multiple levels of conceptual abstraction has been well known since classical times. Aristotle first provided a compelling taxonomic account of both in his Poetics (see Hutton, 1982, for a translation), and computationalists have been fascinated by this perspective ever since. While the core idea has survived relatively unchanged, one must discriminate theories that apparently presume a static type-hierarchy to be sufficient for all abstraction purposes (e.g., see Fass, 1988), from theories that posit the need for a dynamic type hierarchy (e.g., Way, 1991; Veale, 2003). One must also differentiate theories that have actually been implemented (e.g., Fass, 1988; Veale, 2003, 2004) from those that are either notional or that seem to court computational intractability (e.g., see Way, 1991). Perhaps most meaningfully, one must differentiate theories and implementations that assume hand-crafted, purpose-built ontologies (e.g., Fass, 1988) from those that exploit an existing large-scale resource like WordNet (e.g., Veale 2003,2004). In the former, one has the flexibility to support as many functional abstractions like *cutting-agent* as are believed necessary, but at the cost of appearing to anticipate future analogies by hand-crafting them into the system.

This current work follows the latter course. We intend to automatically construct a new taxonomy of analogically-useful abstractions like *cutting-agent*, as a side-effect of grafting the constructive semantic content of HowNet onto the differential backbone of WordNet.

3 Comparing WordNet and HowNet

Generalization can be considered “controlled” if, when moving to a higher level of abstraction in a taxonomy, a conceptual system is able to precisely quantify that meaning which is lost. In this sense at least, most large-scale taxonomies do not provide a significant degree of control. Perhaps nowhere is this observation more keenly felt than in weak lexical ontologies like Princeton WordNet (PWN). In PWN (Miller, 1995), generalization of a concept/synset does not generally yield a functional or behavioural abstraction of the original concept. WordNet’s taxonomy is designed not to capture common causality, function and behaviour, but to show how existing lexemes relate to each other. For example, the common abstraction that unites *{surgeon, sawbones}* and *{tree_surgeon}* is not a concept that captures a shared sense of repair, improvement or care, but *{person, human}*. To be fair, much the same must be said of other taxonomies, such as that of HowNet (Dong, 1988), a Chinese/English semantic dictionary, and even Cyc (see Lenat and Guha, 1990). However, as we shall demonstrate, HowNet contains the necessary basis for such abstractions in its relational semantic definitions.

PWN and HowNet have each been designed according a different theory of semantic organization. PWN is *differential* is nature: rather than attempting to express the meaning of a word explicitly, PWN instead differentiates words with different meanings by placing them in different synsets, and further differentiates synsets from one another by assigning them to different positions in its ontology. In contrast, HowNet is *constructive* in nature, exploiting sememes from a less discriminating taxonomy than PWN's to compose a semantic representation of meaning for each word sense.

Nonetheless, HowNet compensates strongly with its constructive semantics. For example, HowNet assigns the concept *surgeon*|医生 the following definition:

{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 the agent of a doctoring activity." The *{~}* serves as a self-reference here, to mark the location of the concept being defined in the given semantic structure. The oblique reference offered by the tilde construct serves to make the definition more generic (thereby facilitating analogy), so that many different concepts can conceivably employ the same definition. Thus, HowNet uses the above definition not only for surgeon, but for medical workers in general, from orderlies to nurses to internists and neurologists.

3.1 Why Combine Them?

The pros and cons of WordNet and HowNet are nicely complementary: the differential organization of WordNet affords a rich (if not functionally oriented) taxonomy that can be used to assess semantic similarity between potential analogues (see Veale, 2003); in contrast, HowNet's constructive semantics provide the causal basis for identifying analogical candidates for WordNet to assess. However, rather than attempt to utilize both HowNet and WordNet simultaneously, a more practical yet elegant solution can be had if the advantages of one are transplanted to the other. Since analogical systems already exist for WordNet (e.g., Veale, 2004), it thus makes sense to convert the constructive knowledge of HowNet into a differential form that WordNet can readily import.

4 Extracting Functional Structure

Our scheme for converting HowNet's constructive definitions into a more differential form hinges on the use of the tilde as a self-reference in relational structures. For instance, consider the semantic definition that HowNet gives to *repairman|修理工:*

{human|人:HostOf={Occupation|职位},
{repair|修理:agent={~}}}

Noting the position of *{~}*, we can infer that a repairman is the agent of a repairing activity, or in differential terms, a *repair-agent*. Now, since HowNet defines *repair|修理* as a specialization of the reinstatement activity *resume*|恢复, we can further establish *repair-agent* as a specialization of *resume-agent*.

resume-agent
- ***repair-agent***
  - repairman|修理工
  - watchmaker|钟表匠
- ***doctor-agent***
  - surgeon|医生
  - herbalist|药农
- ***amend-agent***
  - reviser|修订者

Figure 1. Portion of a three-level functional hierarchy derived from HowNet.

This double layer of abstraction establishes a new taxonomy that organizes word-concepts according to their analogical potential, rather than their formal ontological properties. For instance, as shown in Figure 1, *resume-agent* encompasses not only *repair-agent*, but *doctor-agent*, since HowNet also defines the predicate *doctor|医治* as a specialization of the predicate *resume*|恢复.

In general, given a semantic fragment *F:role={~}* in a HowNet definition, we create the new abstractions *F-role* and *F'-role*, where *F'* is the immediate hypernym of *F*. The role in question might be *agent*, *instrument*, *location*, *patient*, or any other role that HowNet supports. By way of example, Figure 2 below displays a partial hierarchy derived from the HowNet semantics of various form-altering tools.

AlterForm-instrument

cut-instrument

razor|剃刀

knife|刀

stab-instrument

lance|长矛

sword|宝剑

split-instrument

grater|擦菜板

glasscutter|玻璃刀

break-instrument

scissors|剪

chainsaw|油锯

Figure 2. a hierarchy of form-altering instruments derived from instances of AlterForm|变形状

5 Transplanting Functional Structure

The new, functional taxonomy that results from this extraction of relational structure is differential in nature, but is still HowNet-specific inasmuch as it comprises HowNet senses. To actually graft this hierarchy onto WordNet, we need to first map the relevant HowNet senses to their WordNet equivalents.

We employ a multi-heuristic approach to alignment, in which a variety of symbolic strategies are used to suggest candidate mappings between HowNet and WordNet senses. Some key strategies are the following:

Synonymy:

If the Chinese character-string *xyz* translates as the English words *a, b* in HowNet, and *s = {a, b,…}* is a synset in WordNet, then align *xyz* → *s*. E.g., 长 translates as "head" and "chief" in HowNet, and thus should be aligned with the synset *{chief, head}*.

Generalization:

If the Chinese character-string *xyz* translates as the English words *a* and *b* in HowNet, and there exists a synset *s1 = {a, …}* in WordNet that is a hyponym of a synset *s2 = {b, …}*, then align *xyz* → *s1, s2*. E.g., 手术刀 translates as both "scalpel" and "surgical knife" in HowNet while in WordNet *{scalpel}* is a hyponym of *{surgical_knife}*.

Polysemy:

If the Chinese character-string *xyz* translates as the English words *a* and *b* in HowNet, and there exists a synset *s1 = {a, …}* in WordNet whose textual gloss contains "b", then align *xyz* → *s1*. E.g., 圈 translates as both "enclosure" and "pen" in HowNet, while in WordNet the gloss of *{pen}* is "an *enclosure* for confining livestock".

Compounding:

If the Chinese character-string *xyz* translates as the English word *a*, and the suffix string *yz* translates as the English word *b* such that a synset *s1 = {a, …}* is a hyponym of a synset *s2 = {b, …}* in WordNet, then align *xyz* → *s1* and align *yz* → *s2*. E.g., 语言学家 translates as "linguist" in HowNet while 学家 translates as "scientist", and of the two senses of "linguist" in WordNet, only one is a hyponym of *{scientist}*.

For an alternative, statistical approach, the reader is directed to Carpuat *et al.* 2002.

5.1 Assessing Transplant Quality

To assess the effectiveness of the alignment strategies of section 5, we enlisted a bilingual human judge – fluent in both English and Chinese – to hand-align 5000 Chinese noun entries chosen at random from HowNet. Of these, almost 1500 words comprise two-characters while the remainder comprise a single-character each. Table 1 presents the precision/recall for each strategy on all words, while Table 2 separately presents precision and recall results for multi-character words only (e.g., 语言学家 but not 长).

Table 1. Precision / recall for all 5000 Chinese words.

Heuristic	Precision *all words*	Recall *all words*
Synonymy	0.79	0.12
Generalization	0.73	0.09
Polysemy	0.66	0.11
Compounding	0.63	0.13
Mod Compounding	0.61	0.11
Head Compounding	0.87	0.04
Any one of above	0.64	0.47
Any two of above	0.74	0.17

Table 2. Precision / Recall for multi-char Chinese words.

Heuristic	Precision *multi-char words*	Recall *multi-char words*
Synonymy	0.96	0.23
Generalization	0.88	015
Polysemy	0.74	0.17
Compounding	0.64	0.44
Mod Compounding	0.62	0.41
Head Compounding	0.87	0.18
Any one of above	0.64	0.56
Any two of above	0.89	0.24

Tables 3 and 4 present the results of a corresponding experiment on a hand-annotated gold-set of 12,000 Korean-to-English sense mappings. Korean is, by most estimates, at least 60% Sinitic in origin. Interestingly, this larger test set, the bulk of which are mult-character words, yields better results than the smaller Chinese test-set.

Table 3. Precision / recall for 12,000 Korean words.

Heuristic	Precision *all words*	Recall *all words*
Synonymy	0.90	0.28
Generalization	0.91	0.23
Polysemy	0.71	0.23
Compounding	0.73	0.27
Mod Compounding	0.69	0.18
Head Compounding	0.80	0.15
Any two of above	0.83	0.36
Any one of above	0.73	0.62

Table 4. Precision / recall for multichar Korean words.

Heuristic	Precision *Multi-char words*	Recall *Multi-char words*
Synonymy	0.90	0.30
Generalization	0.91	0.23
Polysemy	0.71	0.24
Compounding	0.73	0.31
Mod Compounding	0.70	0.21
Head Compounding	0.80	0.16
Any two of above	0.83	0.37
Any one of above	0.74	0.63

If we limit our an alignment heuristics to multi-character words only, as reported in Tables 2 and 4, we notice a marked improvement in precision/recall across all heuristics. We believe this is a consequence of the semantic transparency offered by Chinese and Sino-Korean orthography, in which each character is often a distinct meaning-carrying morpheme. Because multi-character words are more semantically constrained by their compound form, they are much less ambiguous on average, and thus less troublesome to map. Indeed, since words of one character form a closed set, these results suggest that we should employ hand alignment for mono-character words while employing automatic alignment for the much larger set of compound words.

6 Evaluating Analogical Competence

We evaluate the analogical potential of the transplanted taxonomy using four criteria: *topology* – the branching structure of the new taxonomy dictates its ability to generate analogies; *coverage* – the percentage of unique HowNet definitions that can be functionally re-indexed in the new taxonomy; *recall* – the percentage of unique definitions for which at least one analogy can be found using the new taxonomy; and *parsimony*– the percentage of abstractions in the new taxonomy that can be used to generate analogies.

6.1 Topological Characteristics

The new functional taxonomy contains 1579 mid-level abstractions and 838 upper-level abstractions. In total, the taxonomy contains only 2219 unique abstractions, revealing that in

8% of cases, the upper-level abstraction of one concept serves as the upper-level abstraction of another.

Analogies will be generated only if two or more unique concept definitions are co-indexed under the same mid-level or upper-level abstraction in the new functional taxonomy. For example, knight|骑士 and gladiator.|打斗者 are both co-indexed directly under the mid-level abstraction *fight-agent*. Likewise, gladiator|打斗者 is indexed under *HaveContest-agent* via *fight-agent*, while footballer|足球运动员 is indexed under *HaveContest-agent* via *compete-agent*. The upper-level of abstraction, represented here by *HaveContest-agent*, is necessary to facilitate analogy between semantically distant concepts.

Nonetheless, we note that a certain degree of *metaphoric licence* has already been exercised by HowNet's designers in assigning semantic structures, so that even semantically distant concepts can still share the same mid-level abstraction. Creative analogies like "Death is an assassin" can, as shown in Figure 3, be resolved via a single generalization step, while analogies like "Death is a man-eater" require two generalization steps.

MakeBad-agent

  kill-agent

    assassin|刺客

    Death|死神

  attack-agent

    intruder|侵略者

    man-eater|食人鲨

Figure 3. semantic diversity among concepts with the same mid-level abstraction.

In addition, the tendency toward under-specification in HowNet means that many analogical concepts will already share the same propositional definition. Because HowNet contains 95,407 unique lexical concepts (excluding synonyms) but only 23,507 unique semantic definitions, these definitions must be under-specified to the extent that many are shared by non-identical concepts. For instance, *cart*|板车 and *bicycle*|单车 are all simply defined as manual vehicles in HowNet.

6.2 Analogical Coverage

Since the new taxonomy is derived from the use of *{~}* in HowNet definitions, both the coverage and recall of analogy generation crucially depend on the widespread use of this reflexive construct. However, of the 23,507 unique definitions in HowNet, just 6430 employ this form of self-reference. The coverage of the new taxonomy is thus 27% of HowNet definitions.

6.3 Analogical Recall

A majority of the abstractions in the new taxonomy, 59%, serve to co-index two or more HowNet definitions. Overall, analogies are generated for 6184 unique HowNet definitions, though these individual definitions may have many different lexical realizations. The recall rate thus is 26% of HowNet's 23,507 unique definitions, or 96% of the 6430 HowNet definitions that make use of {~}. The most productive abstraction is *control_agent*, which serves to co-index 210 unique definitions.

6.4 Parsimony of Abstraction

Overall, 1,315 of the 2219 nodes in the new taxonomy prove useful in co-indexing two or more unique definitions, while 904 nodes serve to index just a single definition. The parsimony of the new taxonomy is thus 59%, which reveals a reasonable, if not ideal, level of representational uniformity across HowNet's semantic definitions.

7 Acknowledgements

While just 27% of HowNet's definitions are sufficiently structured to support analogy, we are encouraged that almost all of this generative potential can be achieved with a new functional taxonomy that is straightforward and efficient to construct. Furthermore, it is possible to graft this new taxonomy directly onto WordNet without doing violence to WordNet's differential form. We are confident then that a full synergy of

WordNet and HowNet will be significantly richer, for analogical purposes, than the sum of the individual parts.

Nonetheless, we have merely scratched the surface of what can usefully be grafted from HowNet onto WordNet. For instance, the full semantic richness of Chinese orthography has yet to be mined for analogical utility. Most Chinese entries in HowNet are multi-character terms whose composite orthography affords a kind of semantic transparency that other writing systems (like that of English) do not possess. Thus, 手术刀, meaning "scalpel", is a composition of characters *and* ideas, since 手术 means "surgery" and 刀 means "knife". Likewise, 哲学家 ("philosopher"), is a composition of 哲学 ("philosophy") and 家 ("specialist"). In turn, philosophy|哲学 is organized by HowNet as a specialization of knowledge|知识, as are the concepts logic|辩学, mathematics|数学, lexicography|词典学 and even midwifery|产科学. By decomposing compound terms and generalizing the extracted modifiers, yet another three-level taxonomy can be constructed. From these examples the partial taxonomy of Figure 4 can be derived.

Knowledge-instrument

  Mathematics-human

    mathematician|数学家

  Philosophy-human

    philosopher|哲学家

  Midwifery-human

    midwife|产科

Figure 4. Portion of an alternate three-level hierarchy derived from compound Chinese terms.

The analogical potential of this ontologization becomes clear when one notices that it supports the classical analogy of philosopher as midwife.

References

Carpuat, M., Ngai, G., Fung, P., and Church, K.W. (2002). Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet. *The Proceedings of GWC 2002, the 1st Global WordNet conference*, Mysore, India.

Dong, Z. (1988). Knowledge Description: What, How and Who? *The Proceedings of the International Symposium on Electronic Dictionaries*, Tokyo, Japan.

Falkenhainer, B., Forbus, K. and Gentner, D. (1989). Structure-Mapping Engine: Algorithm and Examples. *Artificial Intelligence*, 41, pages 1-63.

Fass, D. (1988). An Account of Coherence, Semantic Relations, Metonymy, and Lexical Ambiguity Resolution. In: Small, S. I, Cottrell, G. W., Tanenhaus, M.K. (eds.): *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology and Artificial Intelligence*. Morgan Kaufman, San Mateo California.

Gan, Kok-Wee and Wong, Ping-Wai. (2000). Annotating information structures in Chinese text using HowNet. *The proceedings of the 2nd Chinese Language Processing Workshop, Association for Computational Linguistics 2000 Conference*, October 2000, Hong Kong, pp 85-92.

Hutton, J. (1982). *Aristotle's Poetics*. Norton, New York.

Lenat, D., Guha, R.V. (1990). Building Large Knowledge-Based Systems. Addison Wesley.

Miller, G. A. (1995). WordNet: A Lexical Database for English. Communications of the ACM, Vol. 38 No. 11.

Veale, T., Keane, M. T. (1997). The Competence of Sub-Optimal Structure Mapping on 'Hard' Analogies. *The proceedings of IJCAI'97, the International Joint Conference on Artificial Intelligence*, Nagoya, Japan. Morgan Kaufman, San Mateo California.

Veale. T. (2003). Dynamic Type Creation in Metaphor Interpretation and Analogical Reasoning: A Case-Study with WordNet. *The proceedings of ICCS2003, the 2003 International Conference on Conceptual Structures*, Dresden, Germany.

Veale, T. (2004). WordNet sits the S.A.T.: A Knowledge-Based Approach to Lexical Analogy. *The proceedings of ECAI'2004, the 16th European Conf. on Artificial Intelligence*. John Wiley: London.

Way, E. C. (1991). *Knowledge Representation and Metaphor. Studies in Cognitive systems*. Kluwer Academic Publishers.

Wong, S.H.S. (2004). Fighting Arbitrariness in WordNet-like Lexical Databases – A Natural Language Motivated Remedy. *The proceedings of GWC 2004, the 2nd Global WordNet conference.* Edited by Sojka, Pala, Smrz, Fellbaum, Vossen.