WordNet sits the S.A.T.
A Knowledge-Based Approach to Lexical Analogy

Tony Veale[1]

Abstract. One can measure the extent to which a knowledge-base enables intelligent or creative behavior by determining how useful such a knowledge-base is to the solution of standard psychometric or scholastic tests. In this paper we consider the utility of WordNet, a comprehensive lexical knowledge-base of English word meanings, to the solution of S.A.T. analogies. We propose that such analogies test a student's ability to recognize and estimate a measure of pairwise analogical similarity, and describe an algorithmic formulation of this measure that uses the taxonomic structure of WordNet. We report that the knowledge-based approach yields a precision at least equal to that of statistical machine-learning approaches.

1 INTRODUCTION

The scholastic aptitude test, or SAT, has provided a standardized means of evaluating applicants to the U.S. college system for decades. One of the most fearsome components of this test, for native and non-native speakers of English alike, is a collection of lexical analogies whose solutions require knowledge of much more than just vocabulary. SAT analogies also require an understanding of the subtle relationships that constitute world knowledge. The following are typical examples:

Example 1. a. Ostrich is to Bird as:
b. Cub is to Bear
c. Lion is to Cat
d. Ewe is to Sheep
e. Turkey is to Chicken
f. Jeep is to Truck

Example 2. a. Courier is to Message as:
b. Soldier is to Battle
c. Student is to Knowledge
d. Judge is to Judgment
e. Prophet is to God
f. Athlete is to Prowess

A SAT analogy comprises a source-pairing of concepts/terms (e.g., Ostrich and Bird) and a choice of (usually five) possible target pairings, only one of which accurately mirrors the source relationship. The others typically share a literal or thematic similarity with the source concepts and thus serve as distractors that can lure the student into making a false analogy. The difficulty level of a SAT analogy is a function of both the subtlety of the source relationship and the similarity of the distractors. In the example above, the source relationship can be glossed as "A is a *big* type of B", making (c) the required answer (in contrast, (b) and (d) both suggest a relationship "B is a big type of A"). Analogies offer a progressive means of testing a student's understanding of a domain because they do not encourage rote learning. In fact, because analogies are typically used to communicate salient domain insights, the understanding that an analogy is intended to test can be acquired by the student from the analogy itself at solution time.

Our goal in this research is to measure the effectiveness of WordNet [1] – a comprehensive lexical ontology for English – as a knowledge-base for a SAT-solver. There are both theoretical and practical dimensions to this goal. From a completeness perspective, we want to determine whether WordNet has the subtlety of representation required to solve problems that demand insight and intelligence from human students. While we do not fully subscribe to the psychometric view of AI offered by [2], we believe there is much to be said for the idea of subjecting AI systems and their knowledge-bases to the same tests that are used to quantify, however imperfectly, intelligence in humans, if we are to properly appreciate the true potential of these systems. From a practical perspective, we see this research as providing a basis for building autonomous software tutors that are capable of inventing, posing and grading their own test problems for human students, while also being able to solve these problems themselves from first principles.

2 RESEARCH CONTEXT

Analogy is a much studied aspect of human intelligence that suggests a variety of theoretical approaches, though perhaps none are as well known as the structure-mapping school of [3]. Structure-mapping is essentially a domain-neutral approach that relies of the structural richness and systematicity of a knowledge-representation to yield a mapping based on graph isomorphism. The approach works best when two domains are rich in the higher-order causal structures that can guide the mapping process to a single best mapping, but it can flounder when these structures are missing or flat (see [4]). The latter is true of WordNet, which contains some non-taxonomic structure (such as part-whole relations) but nothing of the kind that would allow a structure-mapping approach to flourish.

An alternate approach that specializes in lexical analogy is offered by taxonomy-driven systems like [5] and [6]. In this situation, where the most important kind of relationship is that encoded by *isa* links, WordNet can certainly support competent analogical reasoning about taxonomic analogies

[1] Department of Computer Science, University College Dublin, Belfield, Dublin, Ireland.

and metaphors, though not, it must be said, without the aid of various parsing techniques that allow feature information stored in free-text annotations to be extracted [6].

Some success has been reported regarding the specific problem of solving SAT analogies. [7] describe a heterogeneous solution framework which brings to bear different approaches and knowledge-sources on the problem, integrating these distinct modules using a weighting system that can be optimized using machine learning. One such module employs a knowledge-based approach based on WordNet, but this module does not significantly out-perform random choice. Another employs web-search as a means of determining the different ways that the terms in a candidate pairing can relate to each other at a syntactic level. This latter method is simple but ingenious. Though the relations that link term pairs in an analogy are unknown, web search using a set of queries incorporating each term pairing can be used to construct a vector of underspecified textual features that are symptomatic of the corresponding relation. These feature-queries are essentially vague templates for connecting two terms, like X-is-Y, Y-of-X, etc., and are underspecified enough to transcend domain boundaries. These vectors, once calculated, can then be compared using a cosine measure that should correlate with the similarity between the corresponding relations. This feature vector approach yields some success on its own, but works best when its output is integrated with that of other modules in a weighted voting scheme.

In our current research, we wish to see if WordNet can be made to deliver equivalent analogical competence to information-based approaches. This would be significant result for WordNet as a knowledge repository, and for SAT solving in general, since we should expect knowledge-based approaches to offer a more transparent and extensible model.

3 THE KNOW-BEST APPROACH

Since analogy is a relational phenomenon, the ideal knowledge-based approach to SAT analogies should also be a relational one. In such an approach, the relation implicit in each pairing of terms would first be determined, so that the target pairing that suggested the same relation as the source pairing could then be identified as the best mapping. However, the relative scarcity of relational structure in WordNet, the same lack that also makes structure-mapping unsuitable, makes the determination of these relationships difficult or impossible. The kinds of relation one finds in WordNet, such as *isa*, *part-of* and *antonym-of*, are simply not the stuff that SAT analogies are made of.

This observation is suggested by an analysis of a large corpus of analogies provided by the authors of [7], and is supported in the analogies of examples 1 and 2. While WordNet represents a throve of *isa* relations, the analogy of example 1 requires us to construct a *big-version-of* relation if the correct answer, (c), is to be distinguished from the close distractors (b) and (d). In some cases, such nuanced relations can be extracted from the textual glosses that annotate sense entries in WordNet (e.g., see [6]), but these glosses lack a consistent form in the lexicon as a whole. Relation extraction is certainly a valid avenue for SAT analogy generation, which has no requirement to be exhaustive, but not for SAT analogy interpretation.

We describe here a knowledge-based approach called KNOW-BEST (KNOWledge-Based Entertainment and Scholastic Testing). In the absence of a robust ability to determine arbitrary relationships among terms, KNOW-BEST employs the more general notion of analogical similarity to rank possible solution candidates. Most theories of analogy give prominence to the notion of similarity (e.g., [3]), of a kind that transcends simple literal similarity. Based on a purely literal reading of similarity, a solver would mistakenly choose (e) in our example analogy, since turkeys and chickens are quite close to ostriches and birds in taxonomic terms. Analogical similarity is a pairwise measure that reflects the fact that it is not just concepts, but the implicit relationships between them, that are being compared. For reasons of KB-completeness we cannot know this relationship, but we can assume that it will be partly determined by the types of concept it is used to relate. The specifics of analogical similarity are the subject of the next section.

Given such a measure, the solution mechanism can be described as follows:

1. The non-noun lexical terms in each pairing are first nominalized, where possible, to permit the underlying concept in the WordNet noun taxonomy to be used instead. Using simple morphology rules, “serene” can be transformed into “serenity” and “tranquil” can be transformed into “tranquility”. The latest version of WordNet, v2.0, has explicit morpho-semantic links between word pairs like these. This transformation is valuable because the noun taxonomy is the most richly developed in WordNet, and the subsumption relations it contains will form a key part of the analogical similarity measure.

2. Each pairing of terms in the analogy undergoes a simple path analysis, to determine whether each pair of terms can be connected in WordNet. A highly constrained wave of spreading activation [8] is used to determine these paths, which are limited to a small number of linkages in length. The goal is to discard those pairings that cannot be connected from further analysis, since these pairings are more likely to be red herrings intended to distract the student from the real answer.

3. If the source pairing involves a subsumption relationship (such as Ostrich:Bird), then all candidate target pairings that do not also involve a subsumption relationship in the same direction are discarded. Subsumption is the only relationship we can reliably identify using WordNet (though partonymy relations are marked in WordNet, coverage is not extensive enough to safely allow target pairings to be discarded).

Each remaining pair of target terms is then measured for analogical similarity with the source pairing. This is a pairwise measure σ that requires four, rather than the conventional two, arguments. The pairing with the highest similarity score is then chosen as the best answer, e.g., $\sigma(lion:cat::ostrich:bird) > \sigma(ewe:sheep::ostrich:bird)$.

We now consider the formulation of σ, our measure of analogical similarity.

4 ANALOGICAL SIMILARITY

There exist a variety of methods for measuring inter-sense similarity in WordNet [9]. Since WordNet is, for the most part, a repository of literal knowledge (e.g., ostriches literally are birds), these methods all tend to measure literal similarity. However, since specific relationships are generally tied to specific types of concepts as arguments, a pairwise measure of taxonomic similarity will nonetheless capture some of the relational similarity that exists between concept pairs. For instance, a *contains* relationship requires that the subject be a container; a *uses* relationship generally requires that the subject be a person and the object be an instrument; and so on. This intuition is supported by a key tenet from metaphor research called the *invariance hypothesis*, which states that the image-schematic structure of the source domain tends to be mapped directly to the image-schematic structure of the target [10]. The image-schemas that a domain is organized around comprise some key ontological categories like CONTAINER, PATH, INSTRUMENT, etc. We therefore construct our measure of analogical similarity around a pairwise measure of taxonomic similarity with an additional measure of the invariance of the mapping as determined relative to these ontological categories. We can use any number of measures of taxonomic similarity (e.g., [8], [11]), but we choose to use (1), since it straightforward to implement and reflects our intuition that as one plunges deeper into the taxonomy, ontological distinctions become finer and differences in similarity thus become smaller.

$$\tau(c_i, c_j) = (2*\delta(p_{ij})) / (\delta(c_i) + \delta(c_j)) \quad (1)$$

Here c_i and c_j denote the concepts to be compared, p_{ij} denotes the lowest common parent of these concepts, and $\delta(c_i)$ denotes the depth of concept c_i in the WordNet taxonomy (where root nodes have depth 0). If we hold the relative depth of c_i, c_j and p_{ij} constant while increasing the depth of all three, then $\tau(c_i, c_j)$ will asymptotically tend toward 1. That is, the finer the ontological difference between c_i and c_j, the greater their perceived similarity.

We now introduce a couple of complications to the definition in (1) to make it less literal and more analogical. First, following [12], we add a measure of textual overlap:

$$\tau(c_i, c_j) = (2*\delta(p_{ij}) + 2^{\omega}) / (\delta(c_i) + \delta(c_j) + 2^{\omega}) \quad (2)$$

The quantity ω is the number of adjectival terms that are shared by the WordNet glosses of c_i and c_j, so that the greater the textual overlap between glosses, the greater the perceived similarity of the corresponding concepts [12]. For instance, abbreviations, abridgements, haiku and dwarves are all the more similar by virtue of each being *short* in their own respective ways. Similarity increases whether the features are used in the same or different senses, the latter constituting the figurative phenomenon of *domain incongruence* (e.g., criminals and steaks are each "tough" but in very different ways). The term 2^{ω} rather than ω is added to both numerator and denominator to allow feature sharing to have a significant impact relative to purely taxonomic concerns, thus ensuring that concepts that are taxonomically distant can still be perceived as analogically similar if they share enough common features.

Secondly, we broaden the idea of a lowest common parent p_{ij}. From a strictly taxonomic perspective, p_{ij} is the most specific common hypernym of two concepts c_i and c_j. For analogical purposes, it is useful to consider two hypernyms p_i and p_j of c_i and c_j as the same (i.e., as p_{ij}) if they have similar lexicalizations. For instance, WordNet defines {seed} as hyponym of {reproductive_structure} and {egg} as a hyponym of {reproductive_cell}. *Reproduction* is thus the unifying theme of the analogy seed:plant::egg:bird. The strict taxonomic similarity between seed and egg is very low, as their lowest common WordNet hypernym is the root node {entity, something}. However, if {reproductive_structure} and {reproductive_cell} are treated as equivalent when determining p_{ij}, their analogical similarity ranks much higher. In general, two hyponyms p_i and p_j can be seen as analogically equivalent if they share a common lexical root or modifier. The analogy embargo:trade::helmet:injury can thus be resolved by recognizing that {embargo} and {helmet}, two very different ontological notions, share a middle ground in *protection*, since {embargo} is a form of {protectionism} and {helmet} a form of {protective_covering} in WordNet. In such cases (which are admittedly rare in WordNet), $\delta(p_{ij})$ is calculated to be the depth of the most specific of p_i and p_j.

Given these considerations, we define the pairwise similarity measure π of a candidate analogy $s_i{:}s_j{::}t_i{:}t_j$ as a weighted sum of individual taxonomic similarities:

$$\pi(s_i{:}s_j, t_i{:}\, t_j) = (\alpha*max(\tau(s_i, t_i), \tau(s_j, t_j)) + \beta*min(\tau(s_i, t_i), \tau(s_j, t_j))) / (\alpha+\beta) \quad (3)$$

We currently choose weights of $\alpha = 2$ and $\beta = 1$, thereby giving twice as much influence to the similarities between pairings than to their dissimilarities.

The invariance hypothesis can be implemented, in part, by choosing the key ontological categories that we wish the analogy mapping to preserve. We choose the general categories CONTAINER, COLLECTION/GROUP, LOCATION, ANIMAL/PERSON, ROLE, SUBSTANCE, and INSTRUMENT. When a candidate mapping $s_i{:}s_j{::}t_i{:}t_j$ violates one of these invariants, e.g., when s_i is a collection and s_j is not, this variance diminishes the analogical similarity measure σ as follows:

$$\sigma(s_i{:}s_j, t_i{:}\, t_j) = \pi(s_i{:}s_j, t_i{:}\, t_j) / 10^{\nu-1} \quad (4)$$

where ν is the number of variances, or violations of the invariance hypothesis, implied by $s_i{:}s_j{::}t_i{:}t_j$.

5 EVALUATION

We tested this approach based on analogical similarity using the WordNet knowledge-base (version 1.6) on an independent corpus of 376 SAT analogies. This corpus, which was kindly provided by the authors of [7], comprises real analogy problems collected from the Web, examination papers and text-books. The results are displayed in Table 1.

Table 1. The analogical-similarity model applied to a test corpus of 376 SAT analogies.

Approach	Coverage	Precision
KNOW-BEST	100% (374)	42%
KNOW-BEST on noun:noun source pairings only	56% (211)	45%
KNOW-BEST on entity:entity source pairings only	24% (94)	53%

As reported in Table 1, WordNet is capable of supporting a reasonable performance on the SAT analogies test, attaining a pass mark of 42% when answering all 376 problems (if one neglects to apply negative marking). If WordNet limits itself to those analogies whose source domain is a pairing of two nouns, this competence raises to 45% but only 56% of the corpus is considered. This statistically-insignificant improvement is surprising, since the taxonomic similarity measure τ is designed to make maximal use of the hierarchical structure found the WordNet nouns taxonomy. However, this intuition is borne out if WordNet limits itself to analogies where the source domain is a pairing of two hyponyms of {entity, something}. In this case, the precision jumps to 54% but at the cost of ignoring 76% of the test corpus.

These results compare well with those of [7], who report 26% precision for a knowledge-based approach using WordNet. However, [7] report higher precision levels for a range of different information-driven modules, each achieving precision levels in the 30-40% range, and 45% for a machine-learning approach that combines the results of these individual modules in a weighted fashion. On a special corpus of 100 analogies [7] reports 55% precision for the heterogeneous, machine-learning approach, which is comparable to the results of the entity:entity-only test set above.

5.1. Measuring the Effect of Latent Similarity

KNOW-BEST and the heterogeneous approach of [7] both achieve a pass-grade because, in their own ways, they each attempt to measure the latent similarity between different term-pairs. Analogy is challenging when our intuitions about surface similarity contradict those about deep similarity, and it is precisely this tension between surface and deep that the distractors in a SAT analogy are designed to stimulate. This suggests that other comparison measures that are honed to capture latent similarities might be equally adept at solving SAT analogies. One such technique is Latent Semantic Analysis, or LSA. Indeed, LSA has already shown a strong ability for solving SAT-style synonymy problems [13], so one might expect analogies to comprise another of its core competences. LSA works by statistically analyzing a large corpus of reading materials to form a very-high dimension semantic-space, which is then reduced to its principle factors to yield a tighter space, of still significant dimensionality. Words from the corpus can then be assigned to particular points in this space, so that words that are similar by virtue of occurring in similar contexts will be situated in the same neighborhood of semantic space. A document of words can also be situated within this space by calculating the centroid of the points assigned to its individual words. LSA thus allows the semantic distance between words or documents to be calculated in terms of spatial distance.

There is a highly suggestive *prima facie* case for considering LSA to be of utility in solving SAT analogies. Given a term pairing like courier:message or judge:judgment, we can expect the centroid of each pairing to pinpoint the area of semantic space where the common relation in both pairings, *to deliver*, is located. Thus, the should expect the LSA similarity between term pairs to correlate well with their analogical similarity.

We tested this hypothesis using two experiments over the test corpus of 374 analogies. In the first, each target pairing of every analogy is treated as a two-word document and compared to the source pairing of the analogy, and the target pairing that achieves the highest LSA score is chosen as the best answer. For example, the document "courier message" is compared with the documents "judge judgment", "prophet prediction" etc. The second experiment follows the same procedure, except that the terms of the source and target pairings are switched, so "courier judgment" is compared with "judge message", "courier prediction" is compared with "prophet message", and so on. It may be that such transverse comparisons are more effective at bringing the analogical similarity between pairings to the fore and reducing the distracting effect of literal similarity. The results of both experiments are reported in Table 2. Note that *term-to-term* and *document-to-document* are the two forms of comparison supported by LSA. Results were obtained from the online implementation of LSA at `lsa.colorado.edu`.

Table 2. Two variants of an LSA solver for SAT analogies.

Comparison	Direct	Transverse
Term-to-term	22%	18%
Document-to-document	24%	18%

The results in Table 1 reveal that latent similarity is not at all the same thing as analogical similarity, at least as conceived of in LSA. Note that 20% is the expected precision of a solver based on simple random choice. We believe that LSA fails because its measure of latent similarity is effectively a measure of literal similarity, albeit one that measures implied as well as explicit similarity. This

failure to recognize deep similarities allows the distractor pairings in each analogy to appear more similar to the source pairing than the best target pairing.

6 CONCLUSIONS

This research suggests that a knowledge-based approach to lexical analogy can achieve as much coverage and precision as a purely information-based approach. This conclusion is significant because the knowledge-base has not been constructed especially for the task, but rather is a general purpose, and for the most part, relationally-bare, ontology of English words. The key is to conceive of the analogy-solving task as one of similarity determination, and to create a flexible model of analogical similarity. Literal similarity does play a role, but a minor one, as evidenced by the poor results obtained by the LSA-based approach. This latter approach is a straw-man, to be sure, but nonetheless suggests that a general model of literal similarity – latent or otherwise – is not enough to solve SAT analogies.

The utility of a heterogeneous environment for resolving lexical analogies has been demonstrated in [7], allowing both knowledge-based and information-based approaches to complement each other. The logical next step then is to evaluate the effect on overall competence when KNOW-BEST is employed as a module in this heterogeneous environment. But more that this, it may be useful to integrate the KNOW-BEST approach not just as a module with its own weighted vote, but as a hypothesis-generator that can explain the data retrieved by more information-based techniques. Information extraction techniques can be used to sift through the results of web-queries to identify possible relations between terms, so that these candidates can be analyzed further in the context of WordNet. Both KNOW-BEST and [7] skirt the issue of actually identifying relationships; if integrated, they may together be able to tackle this problem head-on.

REFERENCES

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

[2] Bringsjord, S., Schimanski, B., What is Artificial Intelligence? Psychometric AI as an Answer. *Proceedings of the 18th International Joint Conference on Artificial Intelligence.* Morgan Kaufmann, San Mateo, CA (2003).

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

[4] Veale, T., Keane, M. T., The competence of structure-mapping on hard analogies. *Proceedings of the 15th International Joint Conference on Artificial Intelligence.* Morgan Kaufmann, San Mateo, CA (2003).

[5] Fass, D. *Processing Metonymy and Metaphor*. Ablex (1997), London.

[6] T. Veale, The Analogical Thesaurus. *Proceedings of the 2003 Conference on Innovative applications of Artificial Intelligence.* Morgan Kaufmann, San Mateo, CA (2003).

[7] P. D. Turney, M. L. Littman, J. Bigham, V. Shnayder, Combining independent modules to solve multiple-choice synonym and analogy problems. *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (2003).

[8] M. R. Quillian, Semantic Memory. *Semantic Information Processing*, ed. M. Minsky. MIT Press, Cambridge (1968).

[9] A. Budanitsky, G. Hirst, Semantic Distance in WordNet: An experimental, application-oriented evaluation of five measures. *Proceedings of the Workshop on WordNet and Other Lexical Resources, North-American chapter of ACL.* Pittsburgh. (2001)

[10] G. Lakoff, *Women, Fire and Dangerous Things.* University of Chicago Press, Chicago (1987)

[11] P. Resnik, Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* 11, pp 95-130, (1999).

[12] M. Lesk, Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of ACM SigDoc Conference, Toronto: ACM*, 24-6.

[13] T. K. Landauer, S. T. Dumais, A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240 (1997).