The Analogical Thesaurus

Tony Veale

Department of Computer,
University College Dublin, Belfield, Dublin, Ireland.

Tony.Veale@UCD.ie

Abstract

Innovative applications often occur at the juncture of radically different domains of research. This paper describes an emerging application, called the analogical thesaurus, that arises at the boundary of two very different domains, the highly applied domain of information retrieval, and the esoteric domain of lexical metaphor interpretation. This application has the potential to not just to improve the utility of conventional electronic thesauri, but to serve as an intelligent mapping component in any system that uses analogical reasoning or case-based reasoning.

1 Introduction

A conventional thesaurus, electronic or otherwise, is designed to improve recall at the cost of precision. Such imprecision often arises because there are few, if any, real synonyms in natural language. This is especially so in English, which opportunistically imbues any near synonyms with different nuances of meaning so that any word substitution will necessarily incur some loss of semantic precision. This issue of semantic precision becomes even more vexing when one considers the analogical purposes to which a thesaurus can be put. For example, suppose one wanted to know the Hindu, Roman or Semitic equivalents of the Greek gods Zeus, Ares and Athena, or to know the Muslim version of the bible, a church or a priest?

Whereas a conventional thesaurus is indexed on a single probe word, analogical queries require both a source and a target term, to permit a mapping between two domains to be constructed. Thus, instead of a simple query like "church" or "bible", one can pose more specific queries like "Muslim church" (mosque), "Hindu bible" (the Vedas), "Celtic Ares" (Morrigan) or "Jewish German" (Yiddish).

Clearly then, semantic precision takes on a very different complexion when analogy is involved. Though "mosque" and "synagogue" are not even near-synonyms, each forms a perfect correspondence with the other in the analogy of a "Muslim synagogue". We should thus carefully differentiate between semantic precision (the basis of synonymy), and analogical precision (the basis of analogy and metaphor).

This paper demonstrates how an analogical thesaurus can be constructed from an existing thesaurus/taxonomy like WordNet (Miller, 1995), by marrying techniques from information retrieval and lexical metaphor interpretation. By viewing analog-retrieval as process of information retrieval (IR) over a semi-structured text archive such as WordNet (which imposes a hierarchical structure onto the unstructured text of dictionary word definitions, or *glosses*), techniques such as query-expansion (Jing and Croft, 1994) can be used to maximize recall. Simultaneously, analogical reasoning techniques can be used to filter the results of IR to ensure that only precise domain counterparts are ever presented to the user. We argue that analogical precision is a more reliable indicator of retrieval utility in a thesaurus than semantic similarity, for while a word may evoke many near-synonyms, a well-structured analogy will often generate a single best mapping (Falkenhainer *et al.*, 1989).

2 Thesaurus Retrieval

We begin by sketching a simple information-retrieval model of how an analogical thesaurus might work, and then critique this model to reveal the extra sophistication that is needed. Remember that our goal to not to retrieve near-synonyms, but sensible analogical correspondences, so pre-existing associative structures like the *synset*s in WordNet will be of little help.

Given a source word s (like "Zeus") and a target domain word t (like "Hindu"), we retrieve all word senses c whose textual glosses contain the word t (such as *{yoga}*, *{fakir}*, etc.) From each candidate c_i on this list, we evaluate the structural similarity of c_i to every sense of s using a taxonomic metric (e.g., Resnik, 1999), such as distance to the lowest common hypernym. For example, *{Zeus}* and *{Varuna}* are

structurally similar to the extent that they share the grandparent *{deity, god}*. Those candidate senses c_i whose similarity to s fall below a certain threshold are rejected, while those that remain are ranked in descending order of similarity and displayed to the user. This straw-man approach works well in many cases, and can answer such queries as "Who is the French Beckett?" (Victor Hugo) and "What is Hebrew German?" (Yiddish). However, there are significant problems that make it largely unusable.

First, the most desirable candidate words may not contain the target word t in their glosses, but may only implicitly relate to t via a hypernym. For instance, while *{Varuna}* is a hyponym of *{Hindu_deity}* in WordNet 1.6, it has a meager gloss, "supreme cosmic deity", in which no explicit reference to any aspect of Hindu culture is made. Secondly, taxonomic discrimination of good candidates from bad is only effective on a coarse-grained basis. For instance, we can rely on the structure of WordNet to recognize that hyponyms of *{deity, god}* (such as *{Aditi}* and *{Avatar})* make better candidates than hyponyms of *{person}* (like *{fakir}* or *{Gurkha}*). But we cannot depend on WordNet to discriminate between *{Varuna}* and *{Aditi}* on a taxonomic basis, since the relevant criterion (supremacy within one's pantheon) is not coded taxonomically in WordNet 1.6, but simply stated in the glosses.

The first problem can be resolved by adapting to the vagaries of word usage in WordNet sense glosses, by using query expansion to cast a wider retrieval net. For example, (Jing and Croft, 1994) use a thesaurus to perform query expansion. The second problem requires that we adapt WordNet itself, to make fine-grained taxonomic discrimination a reality. That is, we need to unlock the implicit structural information contained in WordNet's glosses, and reify this information to the level of taxonomic structure. Thus, *{Zeus}* and *{Varuna}* would become hyponyms of a new taxonomic node, *{supreme_deity}*, such that the presence of this node will introduce greater discrimination into structural similarity metrics.

3 Taxonomic Discrimination

Taxonomic systematicity implies that related or analogous domains should be differentiated in the same ways, so that similarity judgments in each domain can be comparable. But in very large taxonomies, this systematicity is often lacking. For example, in WordNet 1.6, the concept *{alphabet}* is differentiated culturally into *{Greek_alphabet}* and *{Hebrew_alphabet}*, but the concept *{letter, alphabetic_character}* is not similarly differentiated into *{Greek_letter}* and *{Hebrew_letter}*. Rather, every letter of each alphabet, such as *{alpha}* and *{aleph}*, is located under exactly the same hypernym, *{letter, alphabetic_character}*. On structural grounds alone then, each letter is equally similar to every other, no matter what alphabet they belong to (e.g., *alpha* is as similar to *aleph* as it is to *beta*).

An analogical thesaurus would thus be unable to separate good analogues from bad using structural similarity, and in examples such as "Jewish alpha", would return the entire Hebrew alphabet as candidates. To achieve competent analogical mapping then, it is vital that these deficiencies are automatically recognized and repaired. We thus identify an important class of taxonomic support structure for analogies that we dub an "analogical pivot", and show how taxonomies like WordNet, which contain relatively few natural pivots, can be automatically enriched with thousands of new pivots that significantly expand its potential for analogical reasoning. Though we limit our current discussion to the WordNet noun taxonomy, we predict that these techniques are equally applicable to other ontologies, like that of Cyc (Lenat and Guha, 1991).

3.1 Analogical Composition

Consider again the analogical query "Hindu Zeus" and how one might resolve it using WordNet. The goal is to find a counterpart for the source concept Zeus (the supreme deity of the Greek pantheon) in the target domain of Hinduism. In WordNet 1.6, *{Zeus}* is a daughter of *{Greek_deity}*, which is turn is a daughter of *{deity, god}*. Now, because WordNet also defines an entry for *{Hindu_deity}*, it requires just a simple concatenative composition of ideas to determine that the "Hindu Zeus" will be daughter of the node *{Hindu_deity}*. More generally, one finds the lowest parent of the head term ("Zeus") that, when concatenated with the modifier term ("Hindu") or some synonym thereof, yields an existing WordNet concept. In effect, the mapping process uses the pivot to construct a target counterpart of the source concept that significantly narrows the space of possible correspondences.

So the Hindu counterpart of Zeus is not *{Hindu_deity}*, but one of the relatively few daughter nodes of this target-domain differentiation of the pivot *{deity, god}*. Compare this compositional approach with the conventional one of taxonomic reconciliation, due to Aristotle (Hutton, 1982), in which two nodes can be considered analogous if they share a common superordinate. This approach still finds considerable traction in computational models today – e.g., see (Fass, 1998) and (Way, 1991) – but it is easily trivialized: in a well designed taxonomy, any two nodes will always share at least one superordinate (even if it is the root node), and so any two concepts will always be potential analogues in such a system.

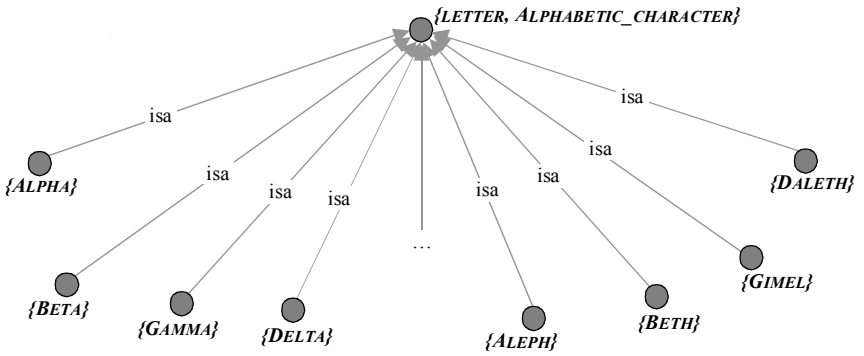

Figure 1: The impoverished sub-taxonomy of {letter, alphabetic_character} *as found in WordNet 1.6*

4 Adding Fine-Grained Distinctions

The current approach avoids triviality by seeking not just any common hypernym, but a hypernym whose lexical form explicitly encodes a juncture of the source and target concepts. However, this compositional approach is still fragile on a number of accounts. First, we note that natural pivots like *{deity, god}* are extremely rare in a taxonomy like WordNet, which has not been explicitly constructed for analogical purposes. For instance, as noted earlier, the WordNet concept *{letter, alphabetic_character}* is not taxonomically differentiated into Greek and Hebrew alphabets, so a mapping cannot be constructed for "Jewish delta" → *{Hebrew_letter}*. Figure 1 illustrates the structure (and lack thereof) of the letter domain in WordNet 1.6. Secondly, even when pivots do exist to facilitate a mapping, what is constructed is a broad target hypernym rather than a specific domain counterpart. One still needs to go from *{Hindu_deity}* to *{Varuna}* (like Zeus a supreme cosmic deity, but of Hinduism), or from *{Hebrew_letter}* to *{daleth}* (like "delta" the fourth letter, but of the Hebrew alphabet).

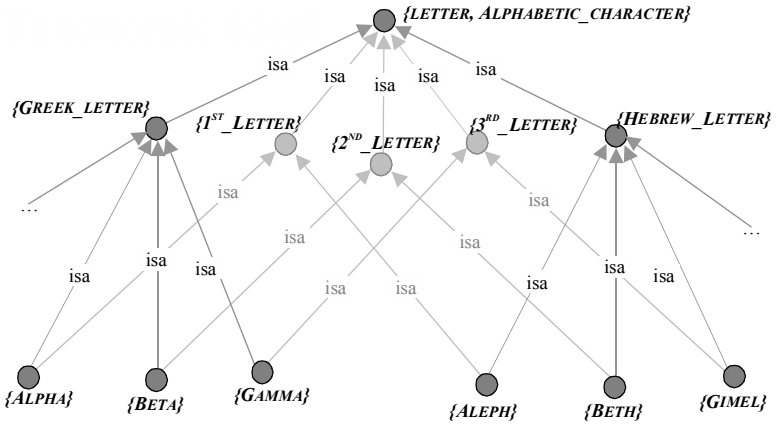

Figure 2: The taxonomic structure of {letter, alphabetic_character} *becomes a richly structured lattice when enriched with a variety of new types like* {Greek_letter} *and* {1st_letter}

However, both problems can be solved by creating additional differentiating nodes that will dissect the taxonomy in new ways. In turn, this will convert the existing hypernyms of these nodes into analogical pivots whose hyponyms are explicitly labeled by domain. For example, the creation of differentiator nodes like *{Greek_letter}* and *{Hebrew_letter}* will transform *{letter, alphabetic_character}* into a pivot that extends into the Greek and Hebrew domains (see Figure 2 below). Nodes such as these act as concatenative junctures of different domains, and serve as signposts from the pivot into more specialized areas of the taxonomy. In contrast, other differentiator nodes may be less ambitious: a new node like *{1st_letter}* will unite just two hyponyms, *{alpha}* and *{aleph}*. However, these lower-level differentiators allow for finer-grained mapping within the established target domain, once the appropriate area of the taxonomy has been identified using the pivot.

Enhancing the differentiating power of WordNet in this way is essentially a task of feature reification. WordNet, like other taxonomies, such as Cyc (Lenat *et al.* 1990), expresses some of its structure explicitly, via *isa*-links, and some of it implicitly, in textual glosses intended for human rather than machine consumption. Fortunately, these glosses are consistent enough to permit automatic extraction of structural features. What is needed then is a means to recognize the word features in these glosses with the most analogical potential, so that they may be 'lifted' to create new taxonomic nodes. We employ two broad criteria to identify the word forms worth reifying in this way:

♦ **Defn**: A lemmatized word-form has *differentiation potential* if it occurs in more than one gloss, but not in too many (e.g., more than 1000). Additionally, there must be a precedent for using the word as an explicit differentiator in at least one existing taxonomic entry.

♦ **Defn**: A lemmatized word-form has *alignment potential* if it can be found in multiple locations of the taxonomy at the same relative depth from a potential pivot.

Consider the word "supreme", which occurs in 42 different WordNet 1.6 glosses, enough to demonstrate cross-domain potential but not too many to suggest vagueness. Additionally, there are three WordNet precedents – *{supreme_court}*, *{supreme_authority}* *and {supreme_being}* – for its explicit use as a differentiator. And of the concepts that "supreme" is used to gloss, six – *{Zeus}*, *{Jove}*, *{Jupiter}*, *{Cronos}, {Wotan}* and *{Varuna}* – are equidistant grand-daughters of the concept *{deity, god}*. The symmetry one expects in analogy is thus present, suggesting that "supreme" can be reified to create a new taxonomic concept *{supreme_deity}*. This situation is illustrated in Figure 3.

In general, any interior non-leaf node of the taxonomy can be a potential pivot node, but from a practical perspective, it makes sense to only consider the atomic concepts that have not already been differentiated. Thus, *{deity, god}* is a potential pivot but *{Greek_deity}* is not, since the latter is already specific to the *{Greek}* domain. For efficiency reasons, we use the following identification rule:

♦ **Defn**: A hypernym X is a potential pivot relative to a hyponym Y if X is the lowest, undifferentiated (atomic) hypernym of Y.

Thus, when we consider the word forms in the gloss of *{Zeus}*, alignability is determined relative to the concept *{deity, god}* rather than *{Greek_deity}*, so that any reification that is performed will create a new differentiation of the former. The analogical thesaurus can exploit a reverse-index of gloss words to the concepts defined using them, to make this identification of alignable features very efficient.

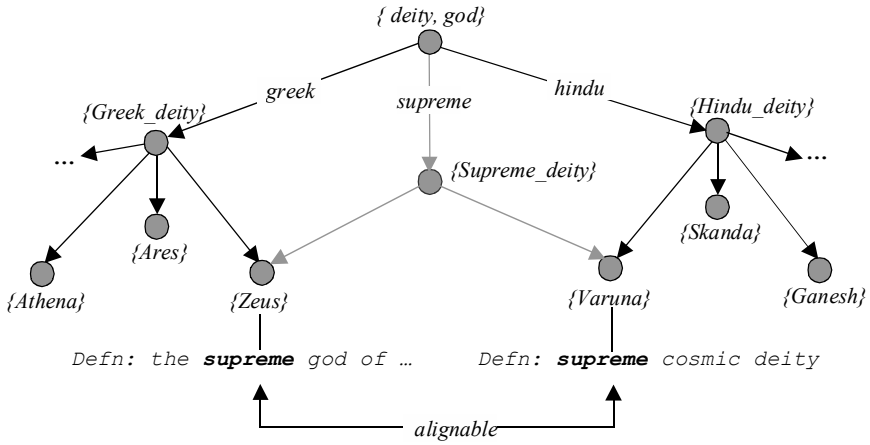

Figure 3: The word-form "supreme" has analogical potential in the gloss of {Zeus}, since it is alignable with another use in {Varuna} at the same relative depth from {deity, god}

5 Experimental Evaluation

To evaluate the effectiveness of an analogical thesaurus based on an automatically enriched version of WordNet, 69,780 unique noun senses in WordNet 1.6 were analyzed, to find those senses which would benefit structurally from the reification of one or more textual features. The glosses of these unique senses collectively contain 35,397 unique (unlemmatized) content words, but because of the strict reification criteria for feature-lifting from glosses, only 2806 of these content words demonstrate both alignability and differentiability. These 2806 words are reified in the

context of specific pivot concepts to add 9822 new differentiator nodes, like *{supreme_deity}* and *{cheese_dish}*, to WordNet. In turn, these nodes serve to differentiate 2737 existing nodes in WordNet, like *{deity}* and *{dish}*, transforming these nodes into analogically-useful pivots.

In total, 18922 noun concepts (27% of the sample) are connected to the new differentiator nodes, via the addition of 28,998 new *isa*-links to WordNet. Each differentiator thus serves to unite an average of 3 daughters apiece. A review of the other 87.2% of differentiators reveals that WordNet is being dissected in new and useful ways, both from the perspective of simple similarity judgments (e.g., the new types achieve a fine-grained clustering of similar ideas) and from the perspective of analogical potential. Overall, the most differentiating feature is "Mexico", which serves to differentiate 34 different pivots (such as *{dish}*, to group together *{taco}, {burrito}* and *{refried_beans}*), while the most differentiated pivot is *{herb, herbaceous_plant}*, which is differentiated into 134 sub-categories (like *{prickly_herb}*). To consider just a few other domains: sports are differentiated into team sports, ball sports, court sports, racket sports and net sports; constellations are divided according to northern and southern hemispheric locales; food dishes are differentiated according to their ingredients, into cheese dishes, meat dishes, chicken dishes, rice dishes, etc.; letters are differentiated both by culture, giving Greek letters and Hebrew letters, and by relative position, so that "alpha" is both a *{1st_letter}* and a *{Greek_letter}*, while "Aleph" becomes both a *{1st_letter}* and a *{Hebrew_letter}*; and deities are further differentiated to yield *{war_deity}, {love_deity}, {wine_deity}, {sea_deity}, {sky_deity}, {thunder_deity}, {fertility_deity}*, and so on.

These new additions, what we might call *dynamic types* since they are created as needed, are primarily intended to increase the precision, rather than the recall rate, of analogical mapping. Consider, for example, the alphabet mapping task, in which the 24 letters of the Greek alphabet are mapped onto the 23 letters of the Hebrew alphabet (as represented in WordNet), and vice versa. The recall rate for the Hebrew to Greek letter task, for both dynamic and static WordNet hierarchies, is 100%, while for the reverse task, Greek to Hebrew, it is 96% (since Greek contains an extra letter). However, the precision of the static hierarchy is only 4%, since every letter of the target alphabet appears equally similar as a candidate mapping (Fig. 2). In contrast, the dynamic hierarchy achieves 96% precision (for Greek to Hebrew alphabets) and 100% (for Hebrew to Greek alphabets).

Table 1 presents a cross-section of the various sub-domains of *{deity, god}* in WordNet as they are organized by new dynamic types such as *{supreme_deity}*. Where a mapping is unavailable for cultural reasons, N/A is used to fill the corresponding cell. In two cases, marked by (*), an adequate mapping could not be generated when one was culturally available. In the case of *{Odin}*, this is due to the brevity of the gloss provided by WordNet 1.6, which defines Odin as a "ruler of the Aesir" rather than the supreme deity of his pantheon. In the case of *{Apollo}*, a Greco-Roman deity, the failure is due to this entity being defined as a Greek deity only in WordNet 1.6.

Common Basis	Greek	Roman	Hindu	Norse	Celtic
supreme	Zeus	Jove	Varuna	Odin *	N/A
wisdom	Athena	Minerva	Ganesh	N/A	Brigit
beauty, love	Aphrodite	Venus	Kama	Freyja	Arianrhod
sea	Poseidon	Neptune	N/A	N/A	Ler
fertility	Dionysus	Ops	N/A	Freyr	Brigit
queen	Hera	Juno	Aditi	Hela	Ana
war	Ares	Mars	Skanda	Tyr	Morrigan
hearth	Hestia	Vesta	Agni	N/A	Brigit
moon	Artemis	Diana	Aditi	N/A	N/A
sun	Apollo	Apollo *	Rahu	N/A	Lug

Table 1: ***Mappings between sub-domains of the type {deity, god} in WordNet 1.6***

The data of Table 1 allows for 20 different mapping tasks in the deities domain (Greek to Roman, Roman to Hindu, etc.). Averaging over the precision and recall achieved on these mapping tasks, we notice a distinct increase in competence when analogical pivots are used to dynamically augment the WordNet hierarchy. Thus, when a dynamic hierarchy is used (i.e., WordNet 1.6 augmented with the pivot identification and creation techniques of section 3), the average recall rate over these 20 mapping tasks is 61%. This moderate performance is as good as one can expect given the nature of the deity systems in these different cultures, since some pantheons are less fleshed out than others. For example, the best precision that a

human can achieve on the Norse to Hindu mapping is just 40%, because of the inherent differences of emphasis in each of these mythological systems. The current system achieves 30% because misses one of the four sensible correspondences, between Odin and Varuna, because of the unhelpful gloss associated with {Odin} in WordNet 1.6.

In contrast for the static hierarchy approach (WordNet 1.6 without additional pivot creation techniques), average recall is significantly lower at 34%, since many concepts are not indexed on the appropriate reference terms due to poorly defined glosses (e.g., Varuna is defined simply as "supreme cosmic deity" in WordNet 1.6, with no explicit reference to Hinduism).

Average precision for the dynamic hierarchy approach is 93.5%, with the loss of 6.5% precision due to the items marked (*) in Table 1. In contrast, average precision for the static hierarchy approach is just 11.5%, and would be lower still if not for the indexing issues arising out of incomplete glosses, which help to reduce the number of incorrect answers that the static hierarchy can actually retrieve.

6 Conclusions

Manually constructed representations on the ambitious scale of WordNet and Cyc are naturally prone to problems of incompleteness and imbalance. The 'one-size-fits-all' nature of the task results in a taxonomy that is often too undifferentiated for precise similarity judgments and too lopsided to support metaphor and analogical mapping. A symptom of this incompleteness is the fact that English glosses or commentaries provide the ultimate level of differentiation, so that one cannot truly differentiate two concepts without first understanding what the glosses mean. This understanding is vital to operation of an analogical thesaurus, so our goal has been to lift implicit discriminators out of the flat text of the glosses and insert them into the taxonomy proper, to facilitate finer similarity judgments and richer analogical mappings.

We view the analogical thesaurus as both a useful component of larger systems (such as those that use analogy for case retrieval) *and* as a useful end-application in its own right, serving to enhance the creative reach of existing thesauri and writer's tools, while simultaneously increasing the precision of these tools. This is perhaps the most counter-intuitive aspect of the analogical thesaurus, as it allows a user to retrieve semantically distant word senses with greater precision than conventional thesauri allow the retrieval of near-synonyms. This paradox of sorts arises out of the distinction between semantic precision and analogical precision: though a more ambitious form of retrieval, analogical retrieval is guided by both a source and target marker to guide its deliberation, allowing one can to meaningfully talk of a perfect correspondence between highly dissimilar words.

References

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

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

Hutton, J. 1982. *Aristotle's Poetics*. Norton: NY.

Jing, Y and Croft, W. B. 1994. An association thesaurus for information retrieval. *The Proceedings of {RIAO}-94, 4th International Conference ``Recherche d'Information Assistee par Ordinateur''*. New York, NY.

Lenat, D. and Guha, R. V. 1990. *Building Large Knowledge-Based Systems*. Addison Wesley: MA.

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Comm. of the ACM*, 38(11).

Resnik, P. 1999. 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 pages 95-130.

Veale, T. and Keane, M. T**.** 1997. The Competence of Sub-Optimal Structure Mapping on Hard Analogies. *The Proc. of IJCAI'97, the International Joint Conference on Artificial Intelligence,* Nagoya, Japan.

Way, E. C. 1991. Knowledge Representation and Metaphor. *Studies in Cognitive systems*. Kluwer Academic: Holland.