Acquiring Naturalistic Concept Descriptions from the Web

Tony Veale, Yanfen Hao
School of Computer Science and Informatics, University College Dublin
Belfield, Dublin 4, Ireland
tony.veale@ucd.ie, yanfen.hao@ucd.ie

Abstract

Many of the beliefs that one uses to reason about everyday entities and events are neither strictly true or even logically consistent. Rather, people appear to rely on a large body of folk knowledge in the form of stereotypical associations, clichés and other kinds of naturalistic descriptions, many of which express views of the world that are second-hand, overly-simplified and, in some cases, non-literal to the point of being poetic. These descriptions pervade our language yet one rarely finds them in authoritative linguistic resources like dictionaries and encyclopaedias. We describe here how such naturalistic descriptions can be harvested from the web in the guise of explicit similes and related text patterns, and empirically demonstrate that these descriptions do broadly capture the way people see the world, at least from the perspective of category organization in an ontology.

1. Introduction

Concepts can meaningfully be described at multiple levels of detail and with varying degrees of literal accuracy, to suit the kind of inferences that are required for a particular application. Simple atomic features, for instance, have long held a practical appeal (e.g., see Katz and Fodor, 1963), since sets of such features can easily be used to discriminate between word meanings in a semantic hierarchy (e.g., see Dong and Dong, 2006). Indeed, when enough features are harvested from the web (see Almuhareb and Poesio, 2004), the words that possess them can be clustered into categories that accurately reflect the structure of a semantic hierarchy like WordNet (Fellbaum, 1998). Furthermore, increasing the detail of these features can yield additional rewards: when features *and* their dimensions are harvested from the web for given words (e.g., hot *and* temperature for "coffee", or hot *and* taste for "chilli"), the accuracy of the semantic hierarchy that can be built via clustering increases also (see Almuhareb and Poesio, 2004, 2005).

The concepts described by these features may themselves be composite structures, and so a given feature may apply to some aspects of a target concept more than others. We say that surgeons are delicate and that poets are sensitive, but it is surely more informative to say that surgeons have delicate hands, or that a poet possesses a sensitive eye. This increased attention to how features are naturally used in everyday linguistic description can pay further dividends when assessing the similarity of two different concepts: for instance, surgeons and artists both have sensitive hands, artists and poets both have sensitive eyes, and poets and orators both have inspiring voices. These descriptions have a naturalistic, almost metaphorical quality that one finds in much of everyday language, in which certain features are communicated by reference to a highly evocative prototype for those features. Thus, a rhinoceros has a thick hide, a lion has a courageous heart, a preacher has an inspiring voice, an eagle has a fierce eye and a statue has a cold visage. These descriptions are "naturalistic" in the sense that they often rely on received linguistic wisdom, stereotypical and even clichéd combinations of ideas that are often literally false. Nonetheless, most English speakers know exactly what is communicated by the description "*the noble soul of a hero*" or "*the cold logic of a computer*".

But, is there truth in metaphors, stereotypes and clichés, or at least enough truth to make this kind of naturalistic description worth harvesting from the web? If so, then these expressions may allow us to acquire a body of semantic features that capture the essence of familiar concepts in a way that a more objective and literal-minded representation can not. We see two ways of testing this hypothesis: if feature sets derived from naturalistic descriptions are more insightful than their more literal counterparts, they should provide a better, more accurate basis for clustering words into semantic hierarchies; or, we should need fewer such features to achieve the same level of clustering accuracy as their less insightful counterparts. In this current work, we demonstrate that this hypothesis is, in fact, true, and that naturalistic descriptions provide a more insightful basis for acquiring and defining semantic features. Our benchmark in this respect is the work of Almuhareb and Poesio (2004, 2005), who demonstrate that large sets of automatically harvested semantic features for given nouns can be used to form a reasonably good semantic hierarchy for those nouns. We argue in this paper that naturalistic descriptions yield comparable semantic clustering ability on the same data sets but with far fewer, and thus more insightful, features.

2. Related Work

Text-based approaches to knowledge acquisition range from the ambitiously comprehensive, in which an entire text or resource is fully parsed and analyzed in depth, to the surgically precise, in which highly-specific text patterns are used to eke out correspondingly specific relationships from a large corpus. Endeavors such as that of Harabagiu *et al*. (1999), in which each of the textual

glosses in WordNet (Fellbaum, 1998) is linguistically analyzed to yield a sense-tagged logical form, is an example of the former approach. In contrast, foundational efforts such as that of Hearst (1992) typify the latter surgical approach, in which one fishes in a large text for word sequences that strongly suggest a particular semantic relationship, such as hypernymy or, in the case of Charniak and Berland (1999), the part-whole relation. Such efforts offer high precision but low recall, and extract just a tiny (but very useful) subset of the semantic content of a text. The KnowItAll system of Etzioni *et al*. (2004) employs the same generic patterns as Hearst (e.g., "*NPs such as NP1, NP2, ...*"), and more besides, to extract a whole range of facts that can be exploited for web-based question-answering. Cimiano and Wenderoth (2007) also use a range of Hearst-like patterns to find text sequences in web-text that are indicative of the lexico-semantic properties of words; in particular, these authors use phrases like "*to * a new NOUN*" and "*the purpose of NOUN is to **" to identify the agentive and telic roles of given nouns, thereby fleshing out the noun's qualia structure as posited by Pustejovsky's (1990) theory of the generative lexicon.

The basic Hearst approach has even proven useful for identifying the meta-properties of concepts in a formal ontology. Völker *et al*. (2005) show that patterns like "*is no longer a|an NOUN*" can identify, with reasonable accuracy, those concepts in an ontology that are not rigid, which is to say, concepts like Teacher and Student whose instances may at any point stop being instances of these concepts. Almuhareb and Poesio (2005) use patterns like "*a|an|the * C is|was*" and "*the * of the C is|was*" to find the actual properties of concepts as they are used in web texts; the former pattern is used to identify value features like *hot*, *red*, *large*, etc., while the latter is used to identify the attribute features that correspond to these values, such as *temperature*, *color* and *size*. Almuhareb and Poesio go on to demonstrate that the values and attributes that are found for word-concepts on the web yield a sufficiently rich representation for these word-concepts to be automatically clustered into a form resembling that assigned by WordNet (see Fellbaum, 1998). Veale and Hao (2007) show that the pattern "*as ADJ as a|an NOUN*" can also be used to identify the feature values associated with a given concept, and argue that because this pattern corresponds to that of the simile frame in English, the adjectival features that are retrieved are much more likely to be highly salient of the noun-concept (the simile vehicle) that is used. Whereas Almuhareb and Poesio succeed in identifying the range of potential attributes and values that may be possessed by a particular concept, Veale and Hao succeed in identifying the generic properties of a concept as it is conceived in its stereotypical form. As noted by the latter authors, this results in a much smaller yet more diagnostic feature set for each concept.

3. The Format of Naturalistic Expressions

As in the approach of Almuhareb and Poesio, we employ textual templates to eke out the features we desire from the text of the web. As described in Veale and Hao (2006), we find the simile template "*as ADJ as a|an NOUN*" particularly effective for identifying the most salient features of the most commonly used vehicles for comparison, returning such instances as "*as hot as an oven*", "*as stiff as a corpse*" and "*as sharp as a knife*". The simile construct is felicitous when the adjective ADJ is a highly salient of NOUN, and when NOUN is a familiar enough term to serve as a vehicle for comparison (though irony subverts this felicity condition). Therefore, the adjectival features that are gathered for a particular noun collectively form a core representation for that noun; thus, for surgeon we obtain *{delicate, sensitive, skilled, professional, clinical, precise, ...}*. Note that these features are highly focused and constitute a picture of the archetypal surgeon. When used directly as a knowledge-representation for inference purposes, these feature sets need to be annotated to remove instances of irony, which we find in almost 15% of instances (intriguingly, when we repeat the experiments for Chinese, on exactly the same large scale, the percentage of ironies is closer to 2%). When used for automatic clustering purposes, no hand-annotation of these features is performed.

Given a set of simple features for a noun, a more detailed picture of how these features interact with a target noun can be acquired using the following template for further web queries: "*the ADJ NOUN of a|an NOUN*", as in "the delicate hands of a surgeon" and "the inspiring voice of a preacher". In conjunction with the first pattern, then, this allows us to acquire a large set of naturalistic <*concept:facet:feature*> triples from the web. The set of facet nouns, such as hands, soul, heart, voice, etc., is limited to the those nouns in WordNet that denote kinds of traits, body parts, qualities, activities and faculties; this allows us to acquire meaningful triples such as <*peacock:tail:colorful*> and <*diamond:beauty:enduring*> while avoiding triples with little or no descriptive insight, such as <*peacock::owner:proud*>. Guarino (1992) suggests that terms like "tail" and "heart" denote *non-relational* attributes, while terms like "temperature" and "size" denote *relational* attributes. One can argue that the term "beauty" can be classified either way, and the current approach seeks out both kinds of terms, which we label here as conceptual *facets*.

These triples can be aligned to specific WordNet senses by exploiting redundancy in the retrieved feature set: when two triples with the same facet and feature (e.g., hands *and* sensitive) are associated with two nouns with senses that are related by synonymy or hypernymy in WordNet (e.g., artist and painter), then those nouns because tied to the specific WordNet senses that manifest this relation. All told, this allows us to associate 18,794 facet:feature tuples with 2032 different WordNet noun senses, for an average of 9 facet:feature pairs per sense. Figure 1 presents two example frame structures that are constructed from these facet:feature pairs, for *peacock* and for *lion*.

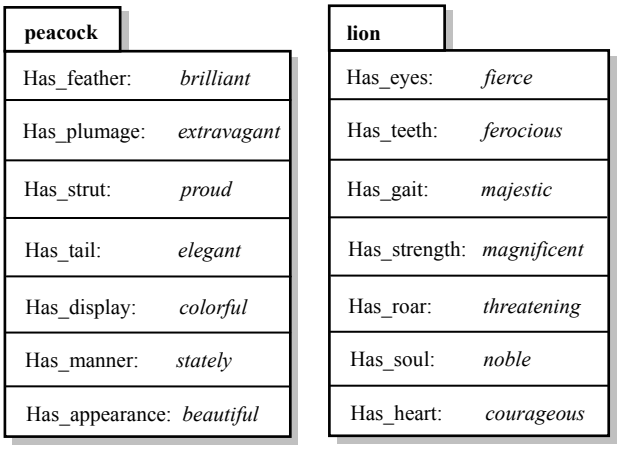

Figure 1: Web-harvested *frame:slot:filler* representations for the concepts Peacock and Lion.

As can be seen from these examples, facet:feature pairs can be seen as slot:filler pairs for purposes of frame construction, and these slot:filler pairs do appear to reflect the most relevant cultural associations for each concept. While one can detect a degree of anthropomorphism and poetic rationalization about these representations, this is a perspective that should be instantly recognizable to native speakers of a language. Nonetheless, it is a perspective that one would be hard pressed to find in a conventional dictionary, except insofar as some lexical concepts may give rise to additional word senses, such as "peacock" for a proud and flashily dressed person.

4. Empirical Evaluation

We report here the results of clustering experiments on the data sets of Almuhareb and Poesio (2004, 2005) which reveal that naturalistic descriptions can achieve the same level of clustering accuracy (when compared to WordNet) as the web-based approach of these earlier authors, but with substantially fewer features. We begin by summarizing the results achieved by Almuhareb and Poesio on their own data-sets, before presenting the results that can be achieved on the same data-set when using naturalistic descriptions.

Almuhareb and Poesio describe two different clustering experiments. In the first, they choose 214 English nouns from 13 of WordNet's upper-level semantic categories, and proceed to harvest property values for these concepts from the web using the pattern "a|an|the * C is|was". This pattern yields a combined total of 51,045 values for all 214 nouns; these values are primarily adjectives, such as hot, black, etc., but noun-modifiers of C are also allowed, such as fruit for cake. They also harvest 8934 attribute nouns, such as temperature and color, using the query pattern "the * of the C is|was". These values and attributes are then used as the basis of a clustering algorithm to partition the 214 nouns into 13 categories, in an attempt to re-construct their original semantic groupings. Comparing these clusters with the original WordNet-based groupings, Almuhareb and Poesio report a cluster accuracy of 71.96% using just values like hot (all 51,045), an accuracy of 64.02% using just attributes like temperature (all 8934), and an accuracy of 85.5% using both values and attributes together (59979 features in total).

In a second, larger experiment, Almuhareb and Poesio select 402 nouns from 21 different semantic classes in WordNet, and proceed to harvest 94,989 property values (again mostly adjectives) and 24,178 attribute nouns from the web using the same retrieval patterns. They then applied the repeated bisections clustering algorithm to this larger data set, and report an initial cluster purity measure of 56.7% using only property values like hot, 65.7% using only attributes like temperature, and 67.7% using both together. Suspecting that noisy features contribute to the perceived drop in performance, those authors then applied a variety of noise filters to reduce the value set to just 51,345 values and the attribute set to just 12,345 attributes, for a size reduction of about 50% in each case. This in turn leads to an improved cluster purity measure of 62.7% using property values only and 70.9% using attributes only. Surprisingly, filtering actually appears to reduce the clustering performance of both sets together to 66.4%.

We replicate here both of these experiments using the same data-sets of 214 and 402 nouns respectively. For fairness, we collect raw descriptions for each of these nouns directly from the web, and use no filtering (manual or otherwise) to remove poor or ill-formed descriptions. Rather, we simply use the pattern "as * as a|an|the C" to collect the top-ranked 100 property values (by web frequency) for every conceptual term C in both data sets. These values are adjectives for the most part, but nouns that serve to modify C are also collected. In all, 2209 raw property values are collected for the 214 nouns of experiment 1, and 5547 raw property values are collected for the 402 nouns of experiment 2. We then use the pattern "the ADJ * of a|an|the C" to collect 4974 attribute nouns for the 214 nouns of experiment 1, and 3952 for the 402 nouns of experiment 2; in each case, ADJ is bound to the raw property values that were acquired using "as * as a|an|the C". A comparison of clustering results is given in Tables 1 and 2.

Approach	***Values only***	***Attr's only***	***All (V + A)***
Almu+Poesio	71.96% (51045 vals)	64.02% (8934 attr)	85.51% (59979 v+a)
Naturalistic Descriptions	70.2% (2209 vals)	78.7% (4974 attr)	90.2% (7183 v+a)

Table 1: Clustering accuracy for experiment 1 (214 nouns)

These tables illustrate that clustering is most effective when it is performed on the basis of both values and attributes (yielding the highest scores, 90.2% and 69.85%, in each experiment respectively).

Approach	*Values only*	*Attr's only*	*All (V + A)*
Almu +Poesio (no filtering)	56.7% (94989 vals)	65.7% (24178 attr)	67.7% (119167 v+a)
Almu.+Poesio (with filtering)	62.7% (51345 vals)	70.9% (12345 attr)	66.4% (63690 v+a)
Naturalistic Descriptions	64.3% (5547 vals)	54.7% (3952 attr)	69.85% (9499 v+a)

Table 2: Clustering accuracy for experiment 2 (402 nouns).

These results thus support the combination of conceptual attributes with specific property values into single naturalistic descriptions which reflect how people actually talk about commonplace concepts. As noted earlier, naturalistic descriptions can be poetic or metaphorical, and may actually express a viewpoint that is overly simplistic, subjective and even technically inaccurate. Nonetheless, these experiments suggest that the linguistic insights we acquire from non-literal descriptions strongly reflect our ontological intuitions about concepts and are more than mere linguistic decorations. Most significantly, we see from these experiments that naturalistic descriptions yield an especially concise representation, since with no filtering of any kind, this approach achieves comparable clustering results with feature sets that are many times smaller than those used in previous work.

5. Conclusions

Naturalistic concept descriptions capture how people actually conceive of and speak about concepts, and as such, they can be markedly different from the objective descriptions conventionally favored by ontologists and semanticists. Naturalistic descriptions often reflect a form of received wisdom that is frequently figurative and sometimes false if judged objectively (e.g., many naturalistic descriptions of animals are based on idealized anthropomorphic models rather than zoological facts). Nonetheless, our results demonstrate that naturalistic descriptions can yield a concise and effective means for organizing knowledge, suggesting that metaphors and similes should be taken very seriously indeed when building linguistic ontologies.

The approach presented here can be used to acquire naturalistic descriptions for languages other than English. To demonstrate this point, we replicated the simile-based feature gathering process for Chinese, a language (and culture) that is distantly removed from English. When the results are hand-annotated, this Chinese experiment yields a simile database of approximately 12,000 unique and well-formed adjective:noun pairings, which is comparable to that achieved for English. We thus find that similes are as widespread in Chinese as they in English, and are thus as commonplace and useful a vehicle for conveying stereotypical knowledge. This demonstrates that harvesting stereotypical knowledge from similes is a workable strategy in both of these languages, and is likely to be just as workable for the other languages (such as Korean) we are about to analyze. Interestingly, the English and Chinese simile sets do not convey precisely the same knowledge, and only 20% or so (or 2440, to be precise) of our English similes are found to have a Chinese translation that can be located on the web. We conclude then by noting that similes seem to be the ideal vehicle for acquiring culturally-specific knowledge from the texts of the web.

6. References

Almuhareb, A. and Poesio, M. (2004). Attribute-Based and Value-Based Clustering: An Evaluation. In *Proc. of EMNLP*. Barcelona, July.

Almuhareb, A. and Poesio, M. (2005). Concept Learning and Categorization from the Web. In *Proc. of the annual meeting of the Cognitive Science society*, Italy, July.

Charniak, E. and Berland, M. (1999). Finding parts in very large corpora. In *Proc. of the 37th Annual Meeting of the ACL*, pp. 57--64.

Dong, Z. and Dong, Q. (2006). HowNet and the Computation of Meaning. World Scientific: Singapore.

Etzioni, O., Kok, S., Soderland, S., Cafarella, M. Popescu, A-M., Weld, D., Downey, D., Shaked, T. and Yates, A. (2004). Web-scale information extraction in KnowItAll (preliminary results). In *Proc. of the 13th WWW Conference*, pp. 100--109.

Fellbaum, C. (ed.). (1998). *WordNet: An electronic lexical database*. The MIT Press. (1985). *A comprehensive grammar of the English*.

Guarino, N. (1992). Concepts, attributes and arbitrary relations: some linguistic and ontological criteria for structuring knowledge base. *Data and Knowledge Engineering*, 8: pp. 249--261.

Harabagiu, S., Miller, G. and Moldovan, D. (1999). WordNet2 - a morphologically and semantically enhanced resource. In *Proc. of SIGLEX-99*, pp. 1--8, University of Maryland.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In Proc. of the 14th Int. Conf. on Computational Linguistics, pp. 539--545, 1992.

Katz, J. J. and Fodor, J. A. (1963). "The structure of a semantic theory." *Language* (2): pp. 170--210.

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics* 17(4), pp. 209--441.

Veale, T. and Hao, Y. (2007). Making WordNet Functional and Context-Sensitive. In *Proceedings of ACL 2007*, Czech Republic, June.

Völker, J., Vrandecic, D. and Sure, Y. (2005). Automatic Evaluation of Ontologies (AEON). In Y. Gil, E. Motta, V. R. Benjamins, M. A. Musen (eds.) *Proc. of the 4th International Semantic Web Conference* (ISWC2005), LNCS 3729:pp. 716--731. Springer Verlag Berlin-Heidelberg.