Multilingual Harvesting of Cross-Cultural Stereotypes

Tony Veale
School of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
tony.veale@ucd.ie

Yanfen Hao
School of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
yanfen.hao@ucd.ie

Guofu Li
School of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
li.guofu.l@gmail.com

Abstract

People rarely articulate explicitly what a native speaker of a language is already assumed to know. So to acquire the stereotypical knowledge that underpins much of what is said in a given culture, one must look to what is implied by language rather than what is overtly stated. Similes are a convenient vehicle for this kind of knowledge, insofar as they mark out the most salient aspects of the most frequently evoked concepts. In this paper we perform a multilingual exploration of the space of common-place similes, by mining a large body of Chinese similes from the web and comparing these to the English similes harvested by Veale and Hao (2007). We demonstrate that while the simile-frame is inherently leaky in both languages, a multilingual analysis allows us to filter much of the noise that otherwise hinders the knowledge extraction process. In doing so, we can also identify a core set of stereotypical descriptions that exist in both languages and accurately map these descriptions onto a multilingual lexical ontology like HowNet. Finally, we demonstrate that conceptual descriptions that are derived from common-place similes are extremely compact and predictive of ontological structure.

1 Introduction

Direct perception of our environment is just one of the ways we can acquire knowledge of the world. Another, more distinctly human approach, is through the comprehension of linguistic descriptions of another person's perceptions and beliefs. Since computers have limited means of human-like perception, the latter approach is also very much suited to the automatic acquisition of world knowledge by a computer (see Hearst, 1992; Charniak and Berland, 1999; Etzioni *et al*., 2004; Völker *et al*., 2005; Almuhareb and Poesio, 2005; Cimiano and Wenderoth, 2007; Veale and Hao, 2007). Thus, by using the web as a distributed text corpus (see Keller *et al*., 2002), a multitude of facts and beliefs can be extracted, for purposes ranging from question-answering to ontology population.

The possible configurations of different concepts can also be learned from how the words denoting these concepts are distributed; thus, a computer can learn that coffee is a beverage that can be served hot or cold, white or black, strong or weak and sweet or bitter (see Almuhareb and Poesio, 2005). But it is difficult to discern from these facts the idealized or stereotypical states of the world, e.g., that one *expects* coffee to be hot and beer to be cold, so that if one spills coffee, we naturally infer the possibilities of scalding and staining without having to be told that the coffee was hot or black; the assumptions of hotness and blackness are just two stereotypical facts about coffee that we readily take for granted. Lenat and Guha (1990) describe these assumed facts as residing in the *white space* of a text, in the body of common-sense assumptions that are rarely articulated as explicit statements. These culturally-shared common-sense beliefs cannot be harvested directly from a single web resource or document set, but must be gleaned indirectly, from telling phrases that are scattered across the many texts of the web.

Veale and Hao (2007) argue that the most pivotal

reference points of this world-view can be detected in common-place similes like "as lazy as a dog", "as fat as a hippo" or "as chaste as a nun". To the extent that this world-view is ingrained in and influenced by how we speak, it can differ from culture to culture and language to language. In English texts, for example, the concept Tortoise is stereotypically associated with the properties *slowness*, *patience* and *wrinkled*, but in Chinese texts, we find that the same animal is a model of *slowness*, *ugliness*, and *nutritional value*. Likewise, because Chinese "wine" has a high alcohol content, the dimension of Strength is much more salient to a Chinese speaker than an English speaker, as reflected in how the word 酒 is used in statements such as 像酒一样浓重, which means "as strong as wine", or literally, "as wine equally strong".

In this paper, we compare the same web-based approach to acquiring stereotypical concept descriptions from text using two very different languages, English and Chinese, to determine the extent to which the same cross-cultural knowledge is unearthed for each. In other words, we treat the web as a large parallel corpus (e.g., see Resnick and Smith, 2003), though not of parallel documents in different languages, but of corresponding translation-equivalent phrases. By seeking translation equivalence between different pieces of textually-derived knowledge, this paper addresses the following questions: if a particular syntagmatic pattern is useful for mining knowledge in English, can its translated form be equally useful for Chinese? To what extent does the knowledge acquired using different source languages overlap, and to what extent is this knowledge language- (and culture-) specific? Given that the syntagmatic patterns used in each language are not wholly unambiguous or immune to noise, to what extent should finding the same beliefs expressed in two different languages increase our confidence in the acquired knowledge? Finally, what representational synergies arise from finding these same facts expressed in two different languages?

Given these goals, the rest of the paper assumes the following structure: in section 2, we summarize related work on syntagmatic approaches to knowledge-acquisition; in section 3, we describe our multilingual efforts in English and Chinese to acquire stereotypical or generic-level facts from the web, by using corresponding translations of the commonplace stereotype-establishing pattern "as ADJ as a NOUN"; and in section 4, we describe how these English and Chinese data-sets can be unified using the bilingual ontology HowNet (Dong and Dong, 2006). This mapping allows us to determine the meaning overlap in both data sets, the amount of noise in each data set, and the degree to which this noise is reduced when parallel translations can be identified. In section 5 we demonstrate the overall usefulness of stereotype-based knowledge-representation by replicating the clustering experiments of Almuhareb and Poesio (2004, 2005) and showing that stereotype-based representations are both compact and predictive of ontological classification. We conclude the paper with some final remarks in section 6.

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 NP_1, NP_2, ..."), 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 value feature 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. However, because the simile frame is often exploited for ironic purposes in web texts (e.g., "as meaty as a skeleton"), and because irony is so hard to detect, Veale and Hao suggest that the adjective:noun pairings found on the web should be hand-filtered to remove such examples. Given this onerous requirement for hand-filtering, and the unique, culturally-loaded nature of the noise involved, we use the work of Veale and Hao as the basis for the cross-cultural investigation in this paper.

3 Harvesting Knowledge from Similes: English and Chinese

Because similes are containers of culturally-received knowledge, we can reasonably expect the most commonly used similes to vary significantly from language to language, especially when those languages correspond to very different cultures. These similes form part of the linguistic currency of a culture which must be learned by a speaker, and indeed, some remain opaque even to the most educated native speakers. In "A Christmas Carol", for instance, Dickens (1943/1984) questions the meaning of "as dead as a doornail", and notes: "I might have been inclined, myself, to regard a coffin-nail as the deadest piece of ironmongery in the trade. But the wisdom of our ancestors is in the simile".

Notwithstanding the opacity of some instances of the simile form, similes are very revealing about the concepts one most encounters in everyday language. In section 5 we demonstrate that concept descriptions which are harvested from similes are both extremely compact and highly predictive of ontological structure. For now, we turn to the process by which similes can be harvested from the text of the web. In section 3.1 we summarize the efforts of Veale and Hao, whose database of English similes drives part of our current investigation. In section 3.2 we describe how a comparable database of Chinese similes can be harvested from the web.

3.1 Harvesting English Similes

Veale and Hao (2007) use the Google API in conjunction with Princeton WordNet (Fellbaum, 1998) as the basis of their harvesting system. They first extracted a list of antonymous adjectives, such as "hot" or "cold", from WordNet, the intuition being that explicit similes will tend to exploit properties that occupy an exemplary point on a scale. For every adjective ADJ on this list, they then sent the query "*as ADJ as **" to Google and scanned the first 200 snippets returned for different noun values for the wildcard *. The complete set of nouns extracted in this way was then used to drive a sec-

ond harvesting phase, in which the query "*as * as a NOUN*" was used to collect similes that employ different adjectives or which lie beyond the 200-snippet horizon of the original search. Based on this wide-ranging series of core samples (of 200 hits each) from across the web, Veale and Hao report that both phases together yielded 74,704 simile instances (of 42,618 unique types, or unique adjective:noun pairings), relating 3769 different adjectives to 9286 different nouns. As often noted by other authors, such as Völker *et al*. (2005), a pattern-oriented approach to knowledge mining is prone to noise, not least because the patterns used are rarely leak-free (inasmuch as they admit word sequences that do not exhibit the desired relationship), and because these patterns look at small text sequences in isolation from their narrative contexts. Veale and Hao (2007) report that when the above 42,618 simile types are hand-annotated by a native speaker, only 12,259 were judged as non-ironic and meaningful in a null context. In other words, just 29% of the retrieved pairings conform to what one would consider a well-formed and reusable simile that conveys some generic aspect of cultural knowledge. Of those deemed invalid, 2798 unique pairings were tagged as ironic, insofar as they stated precisely the opposite of what is stereotypically believed to be true.

3.2 Harvesting Chinese Similes

To harvest a comparable body of Chinese similes from the web, we also use the Google API, in conjunction with both WordNet and HowNet (Dong and Dong, 2006). HowNet is a bilingual lexical ontology that associates English and Chinese word labels with an underlying set of approximately 100,000 lexical concepts. While each lexical concept is defined using a unique numeric identifier, almost all of HowNet's concepts can be uniquely identified by a pairing of English and Chinese labels. For instance, the word "王八" can mean both Tortoise and Cuckold in Chinese, but the combined label tortoise|王八 uniquely picks out the first sense while cuckold|王八 uniquely picks out the second. Though Chinese has a large number of figurative expressions, the yoking of English to Chinese labels still serves to identify the correct sense in almost every case. For instance, "绿帽子" is another word for Cuckold in Chinese, but it can also translate as "green hat" and "green scarf". Nonetheless, green_hat|绿帽子 uniquely identifies the literal sense of "绿帽子" (a green covering) while green_scarf|绿帽子 and cuckold|绿帽子 both identify the same human sense, the former being a distinctly culture-specific metaphor for cuckolded males (in English, a dispossessed lover "wears the cuckold's horns"; in Chinese, one apparently "wears a green scarf").

We employ the same two-phase design as Veale and Hao: an initial set of Chinese adjectives are extracted from HowNet, with the stipulation that their English translations (as given by HowNet) are also categorized as adjectives in WordNet. We then use the Chinese equivalent of the English simile frame "像* 一样ADJ" (literally, "as-NOUN-equally-ADJ") to retrieve a set of noun values that stereotypically embody these adjectival features. Again, a set of 200 snippets is analyzed for each query, and only those values of the Google * wildcard that HowNet categorizes as nouns are accepted. In a second phase, these nouns are used to create new queries of the form "像Noun一样*" and the resulting Google snippets are now scanned for adjectival values of *.

In all, 25,585 unique Chinese similes (i.e., pairings of an adjective to a noun) are harvested, linking 3080 different Chinese adjectives to 4162 Chinese nouns. When hand-annotated by a native Chinese speaker, the Chinese simile frame reveals itself to be considerably less leaky than the corresponding English frame. Over 58% of these pairings (14,867) are tagged as well-formed and meaningful similes that convey some stereotypical element of world knowledge. The Chinese pattern "像*一样*" is thus almost twice as reliable as the English "as * as a *" pattern. In addition, Chinese speakers exploit the simile frame much less frequently for ironic purposes, since just 185 of the retrieved similes (or 0.7%) are tagged as ironic, compared with ten times as many (or 7%) retrieved English similes. In the next section we consider the extent to which these English and Chinese similes convey the same information.

4 Tagging and Mapping of Similes

In each case, the harvesting processes for English and for Chinese allow us to acquire stereotypi-

cal associations between words, not word senses. Nonetheless, the frequent use of synonymous terms introduces a substantial degree of redundancy in these associations, and this redundancy can be used to perform sense discrimination. In the case of English similes, Veale and Hao (2007) describe how two English similes "as A as N_1" and "as A as N_2" will be mutually disambiguating if N_1 and N_2 are synonyms in WordNet, or if some sense of N_1 is a hypernym or hyponym of some sense of N_2 in WordNet. This heuristic allows Veale and Hao to automatically sense-tag 85%, or 10,378, of the unique similes that are annotated as valid. We apply a similar intuition to the disambiguation of Chinese similes: though HowNet does not support the notion of a synset, different word-senses that have the same meaning will be associated with the same logical definition. Thus, the Chinese word "著名" can translate as "celebrated", "famous", "well-known" and "reputable", but all four of these possible senses, given by celebrated|著名, famous|著名, well-known|著名 and reputable|著名, are associated with the same logical form in HowNet, which defines them as a specialization of ReputationValue|名声值. This allows us to safely identify "著名" with this logical form. Overall, 69% of Chinese similes can have both their adjective and noun assigned to specific HowNet meanings in this way.

4.1 Translation Equivalence Among Similes

Since HowNet represents an integration of English and Chinese lexicons, it can easily be used to connect the English and Chinese data-sets. For while the words used in any given simile are likely to be ambiguous (in the case of one-character Chinese words, highly so), it would seem unlikely that an incorrect translation of a web simile would also be found on the web. This is an intuition that we can now use the annotated data-sets to evaluate.

For every English simile of the form $<A_e$ as $N_e>$, we use HowNet to generate a range of possible Chinese variations $<A_{c0}$ as $N_{c0}>$, $<A_{c1}$ as $N_{c0}>$, $<A_{c0}$ as $N_{c1}>$, $<A_{c1}$ as $N_{c1}>$, ... by using the HowNet lexical entries $A_e|A_{c0}$, $A_e|A_{c1}$, ..., $N_e|N_{c0}$, $N_e|N_{c1}$, ... as a translation bridge. If the variation $<A_{ci}$ as $N_{cj}>$ is found in the Chinese data-set, then translation equivalence is assumed between $<A_e$ as $N_e>$ and $<A_{ci}$ as $N_{cj}>$; furthermore, $A_e|A_{ci}$ is assumed to be the HowNet sense of the adjectives A_e and A_{ci} while N_{cj} is assumed to be the HowNet sense of the nouns N_e and N_{cj}. Sense-tagging is thus a useful side-effect of simile-mapping with a bilingual lexicon.

Language	*Precision*	*Recall*	*F1*
English	0.76	0.25	0.38
Chinese	0.82	0.27	0.41

Table 1: Automatic filtering of similes using Translation Equivalence.

We attempt to find Chinese translation equivalences for all 42,618 of the English adjective:noun pairings harvested by Veale and Hao; this includes both the 12,259 pairings that were hand-annotated as valid stereotypical facts, and the remaining 30,359 that were dismissed as noisy or ironic. Using HowNet, we can establish equivalences from 4177 English similes to 4867 Chinese similes. In those mapped, we find 3194 English similes and 4019 Chinese similes that were hand-annotated as valid by their respective native-speaker judges. In other words, translation equivalence can be used to separate well-formed stereotypical beliefs from ill-formed or ironic beliefs with approximately 80% precision. The precise situation is summarized in Table 1.

As noted in section 3, just 29% of raw English similes and 58% of raw Chinese similes that are harvested from web-text are judged as valid stereotypical statements by a native-speaking judge. For the task of filtering irony and noise from raw data sets, translation equivalence thus offers good precision but poor recall, since most English similes appear not to have a corresponding Chinese variant on the web. Nonetheless, this heuristic allows us to reliably identify a sizeable body of cross-cultural stereotypes that hold in both languages.

4.1.1 Error Analysis

Noisy propositions may add little but empty content to a representation, but ironic propositions will actively undermine a representation from within, leading to inferences that are not just unlikely, but patently false (as is generally the intention of irony). Since Veale and Hao (2007) annotate their data-

set for irony, this allows us to measure the number of egregious mistakes made when using translation equivalence as a simile filter. Overall, we see that 1% of Chinese similes that are accepted via translation equivalence are ironic, accounting for 9% of all errors made when filtering Chinese similes. Likewise, 1% of the English similes that are accepted are ironic, accounting for 5% of all errors made when filtering English similes.

4.2 Representational Synergies

By mapping WordNet-tagged English similes onto HowNet-tagged Chinese similes, we effectively obtain two representational viewpoints onto the same shared data set. For instance, though HowNet has a much shallower hierarchical organization than WordNet, it compensates by encapsulating the meaning of different word senses using simple logical formulae of semantic primitives, or sememes, that are derived from the meaning of common Chinese characters. WordNet and HowNet thus offer two complementary levels or granularities of generalization that can be exploited as the context demands.

4.2.1 Adjective Organization

Unlike WordNet, HowNet organizes its adjectival senses hierarchically, allowing one to obtain a weaker form of a given description by climbing the hierarchy, or to obtain a stronger form by descending the hierarchy from a particular sense. Thus, one can go up from kaleidoscopic|斑驳陆离 to colored|彩, or down from colored|彩 to any of motley|斑驳, dappled|斑驳, prismatic|斑驳陆离 and even gorgeous|斑斓. Once stereotypical descriptions have been sense-tagged relative to HowNet, they can easily be further enhanced or bleached to suit the context of their use. For example, by allowing a Chinese adjective to denote any of the senses above it or below in the HowNet hierarchy, we can extend the mapping of English to Chinese similes so as to achieve an improved recall of .36 (though we note that this technique reduces the precision of the translation-equivalence heuristic to .75).

As demonstrated by Almuhareb and Poesio (2004), the best conceptual descriptions combine adjectival values with the attributes that they fill. Because adjectival senses hook into HowNet's upper ontology via a series of abstract taxonyms like TasteValue|美丑值, ReputationValue|名声值 and AmountValue|多少值, a taxonym of the form AttributeValue can be identified for every adjective sense in HowNet. For example, the English adjective "beautiful" can denote either beautiful|美, organized by HowNet under BeautyValue|美丑值, or beautiful|婉, organized by HowNet under gracious|雅 which in turn is organized under GraceValue|典雅值. The adjective "beautiful" can therefore specify either the Grace or Beauty attributes of a concept. Once similes have been sense-tagged, we can build up a picture of most salient attributes of our stereotypical concepts. For instance, "peacock" similes yield the following attributes via HowNet: *Beauty*, *Appearance*, *Color*, *Pride*, *Behavior*, *Resplendence*, *Bearing* and *Grace*; likewise "demon" similes yield the following: *Morality*, *Behavior*, *Temperament*, *Ability* and *Competence*.

4.2.2 Orthographic Form

The Chinese data-set lacks counterparts to many similes that one would not think of as culturally-determined, such "as red as a ruby", "as cruel as a tyrant" and "as smelly as a skunk". One significant reason for this kind of omission is not cultural difference, but obviousness: many Chinese words are multi-character gestalts of different ideas (see Packard, 2000), so that these ideas form an explicit part of the orthography of a lexical concept. For instance, using HowNet, we can see that skunk|臭鼬 is actually a gestalt of the concepts smelly|臭 and weasel|鼬, so the simile "as smelly as a skunk" is already somewhat redundant in Chinese (somewhat akin to the English similes "as hot as a hotdog" or "as hard as a hardhat").

Such decomposition can allow us to find those English similes that are already orthographically explicit in Chinese word-forms. We simply look for pairs of HowNet senses of the form Noun|XYZ and Adj|X, where X and XYZ are Chinese words and the simile "as Adj as a|an Noun" is found in the English simile set. When we do so, we find that 648 English similes, from "as meaty as a steak" to "as resonant as a cello", are already fossilized in the orthographic realization of the corresponding Chinese concepts. When fossilized similes are uncovered in this way,

the recall of translation equivalence as a noise filter rises to .29, while its precision rises to .84 (see Table 1)

5 Empirical Evaluation: Simile-derived Representations

Stereotypes persist in language and culture because they are, more often than not, cognitively useful: by emphasizing the most salient aspects of a concept, a stereotype acts as a dense conceptual description that is easily communicated, widely shared, and which supports rapid inference. To demonstrate the usefulness of stereotype-based concept descriptions, we replicate here the clustering experiments of Almuhareb and Poesio (2004, 2005), who in turn demonstrated that conceptual features that are mined from specific textual patterns can be used to construct WordNet-like ontological structures. These authors used different text patterns for mining feature values (like *hot*) and attributes (like *temperature*), and their experiments evaluated the relative effectiveness of each as a means of ontological clustering. Since our focus in this paper is on the harvesting of feature values, we replicate here only their experiments with values.

Almuhareb and Poesio (2004) used as their experimental basis a sampling of 214 English nouns from 13 of WordNet's upper-level semantic categories, and proceeded to harvest adjectival features for these noun-concepts from the web using the textual pattern "[a | an | the] * C [is | was]". This pattern yielded a combined total of 51,045 value features for these 214 nouns, such as *hot*, *black*, etc., which were then used as the basis of a clustering algorithm in an attempt to reconstruct the WordNet classifications for all 214 nouns. Clustering was performed by the CLUTO-2.1 package (Karypis, 2003), which partitioned the 214 nouns in 13 categories on the basis of their 51,045 web-derived features. Comparing these clusters with the original WordNet-based groupings, Almuhareb and Poesio report a clustering accuracy of 71.96%. In a second, larger experiment, Almuhareb and Poesio (2005) sampled 402 nouns from 21 different semantic classes in WordNet, and harvested 94,989 feature values from the web using the same textual pattern. They then applied the repeated bisections clustering algorithm to this larger data set, and report an initial cluster purity measure of 56.7%. Suspecting that a noisy feature set had contributed to the apparent drop in performance, these authors then proceed to apply a variety of noise filters to reduce the set of feature values to 51,345, which in turn leads to an improved cluster purity measure of 62.7%.

Approach	*accuracy*	*features*
Almuhareb + Poesio	71.96%	51,045
Simile-derived stereotypes	70.2%	2,209

Table 2: Results for experiment 1 (214 nouns, 13 WN categories).

Approach	*Cluster purity*	*Cluster entropy*	*features*
Almu. + Poesio (no filtering)	56.7%	38.4%	94,989
Almu. + Poesio (with filtering)	62.7%	33.8%	51345
Simile-derived stereotypes (no filtering)	64.3%	33%	5,547

Table 3: Results for experiment 2 (402 nouns, 21 WN categories).

We replicated both of Almuhareb and Poesio's experiments on the same experimental data-sets (of 214 and 402 nouns respectively), using instead the English simile pattern "as * as a NOUN" to harvest features for these nouns from the web. Note that in keeping with the original experiments, no hand-tagging or filtering of these features is performed, so that every raw match with the simile pattern is used. Overall, we harvest just 2209 feature values for the 214 nouns of experiment 1, and 5547 features for the 402 nouns of experiment 2. A comparison of both sets of results for experiment 1 is shown is Table 2, while a comparison based on experiment 2 is shown is Table 3.

While Almuhareb and Poesio achieve marginally higher clustering on the 214 nouns of experiment 1, they do so by using over 20 times as many features.

In experiment 2, we see a similar ratio of feature quantities before filtering; after some initial filtering, Almuhareb and Poesio reduce their feature set to just under 10 times the size of the simile-derived feature set.

These experiments demonstrate two key points about stereotype-based representations. First, the feature representations do not need to be hand-filtered and noise-free to be effective; we see from the above results that the raw values extracted from the simile pattern prove slightly more effective than filtered feature sets used by Almuhareb and Poesio. Secondly, and perhaps more importantly, stereotype-based representations prove themselves a much more compact means (by factor of 10 to 20 times) of achieving the same clustering goals.

6 Conclusions

Knowledge-acquisition from texts can be a process fraught with complexity: such texts - especially web-based texts - are frequently under-determined and vague; highly ambiguous, both lexically and structurally; and dense with figures of speech, hyperbolae and irony. None of the syntagmatic frames surveyed in section 2, from the "NP such as NP_1, NP_2 ..." pattern of Hearst (1992) and Etzioni *et al.* (2004) to the "no longer NOUN" pattern of Völker *et al.* (2005), are leak-free and immune to noise. Cimiano and Wenderoth (2007) mitigate this problem somewhat by performing part-of-speech analysis on all extracted text sequences, but the problem remains: the surgical, pattern-based approach offers an efficient and targeted means of knowledge-acquisition from corpora because it largely ignores the context in which these patterns occur; yet one requires this context to determine if a given text sequence really is a good exemplar of the semantic relationship that is sought.

In this paper we have described how stereotypical associations between adjectival properties and noun concepts can be mined from similes in web text. When harvested in both English and Chinese, these associations exhibit two kinds of redundancy that can mitigate the problem of noise. The first kind, *within-language* redundancy, allows us to perform sense-tagging of the adjectives and nouns that are used in similes, by exploiting the fact that the same stereotypical association can occur in a variety of synonymous forms. By recognizing synonymy between the elements of different similes, we can thus identify the underlying senses (or WordNet synsets) in these similes. The second kind, *between-language* redundancy, exploits the fact that the same associations can occur in different languages, allowing us to exploit translation-equivalence to pin these associations to particular lexical concepts in a multilingual lexical ontology like HowNet. While between-language redundancy is a limited phenomenon, with just 26% of Veale and Hao's annotated English similes having Chinese translations on the web, this phenomenon does allow us to identify a significant core of shared stereotypical knowledge across these two very different languages.

Overall, our analysis suggests that a comparable number of well-formed Chinese and English similes can be mined from the web (our exploration finds approx. 12,000 unique examples of each). This demonstrates that harvesting stereotypical knowledge from similes is a workable strategy in both languages. Moreover, Chinese simile usage is characterized by two interesting facts that are of some practical import: the simile frame "像NOUN 一样ADJ" is a good deal less leaky and prone to noise than the equivalent English frame, "as ADJ as a NOUN"; and Chinese speakers appear less willing to subvert the stereotypical norms of similes for ironic purposes. Further research is needed to determine whether these observations generalize to other knowledge-mining patterns.

References

A. Almuhareb and M. Poesio. 2004. *Attribute-Based and Value-Based Clustering: An Evaluation*. In proceedings of EMNLP 2004, pp 158–165. Barcelona, Spain.

A. Almuhareb and M. Poesio. 2005. *Concept Learning and Categorization from the Web*. In proceedings of CogSci 2005, the 27th Annual Conference of the Cognitive Science Society. New Jersey: Lawrence Erlbaum.

C. Dickens. 1843/1981. *A Christmas Carol*. Puffin Books, Middlesex, UK.

C. Fellbaum. 1998. *WordNet, an electronic lexical database*. MIT Press.

E. Charniak and M. Berland. 1999. *Finding parts in*

very large corpora. In proceedings of the 37th Annual Meeting of the ACL, pp 57-64.

F. Keller, M. Lapata, and O. Ourioupina. 2002. *Using the web to overcome data sparseness*. In proceedings of EMNLP-02, pp 230-237.

F. Keller, M. Lapata, and O. Ourioupina. 1990. *Building large knowledge-based systems: representation and inference in the Cyc project*. Addison-Wesley.

G. Karypis. 2003. *CLUTO: A clustering toolkit.* University of Minnesota.

J. L. Packard. 2000. *The Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge University Press, UK.

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

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

M. Hearst. 1992. *Automatic acquisition of hyponyms from large text corpora*. In proceedings of the 14th intenatinal conference on Computational Linguistics, pp 539-545.

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

P. Cimiano and J. Wenderoth. 2007. *Automatic Acquisition of Ranked Qualia Structures from the Web*. In proceedings of the 45th Annual Meeting of the ACL, pp 888–895.

P. Resnik and N. A. Smith. 2003. *The Web as a parallel corpus*. Computational Linguistics, 29(3),pp 349-380.

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

T. Veale and Y. Hao. 2007. *Making Lexical Ontologies Functional and Context-Sensitive*. In proceedings of the 45th Annual Meeting of the ACL, pp 57-64.

Z. Dong and Q. Dong. 2006. *HowNet and the Computation of Meaning*. World Scientific: Singapore.