Unlocking the Latent Creativity of Orthographic Structure

Tony Veale and **Shanshan Chen** [1]

Abstract. Different languages tend to represent different cultural and conceptual perspectives on the world. To the originating culture, such lexicalized perspectives may seem entirely conventional and stale, but to another they may well provide fresh and even innovative insights into the meaning and creative uses of words. In this paper we describe how these insights can be mined from the lexical structure of Chinese, a logomorphemic language that exhibits its semantic structure quite openly in its orthographic realization.

1 Introduction

Whether or not one believes in Wittgenstein's observation that the "limits of my language are the limits of my world", it is a truism that different languages represent different perspectives on the world, and these perspectives are most readily visible in how words are used to carve up this world into concepts. The possibility of translation means that all languages describe the same world in relatively interchangeable ways, yet each language reflects a unique cultural bias by allocating individual words to some concepts and not others. The German word "Schadenfreude", for instance, describes a complex emotion that English and French speakers can also understand, but the very fact that German lexicalizes this concept in a single word (whereas English and French require complex descriptions to convey the same idea) says something interesting both about the German language and the cultural perspective of its speakers.

In this respect, the Chinese written language makes an interesting case in point. Most Chinese words are compounds constructed from an aggregation of morphemic characters, and as such, the orthographic form of a Chinese word can be most revealing about its semantic content, in ways that English words are not. For instance, the Chinese word for "scalpel", 手术刀, is an aggregation of 手术, meaning "surgery", and 刀, meaning "knife" or "sword". The English word "scalpel" cannot be decomposed in this way to reveal its meaning. Likewise, the Chinese word for "mathematician", 数学家, is an aggregation of 数学, meaning "mathematics" or "arithmetic", and 家, meaning "specialist", while 数学 can be further dissected to find the morpheme 数, meaning "number". Most concepts in Chinese are thus conveyed by multi-morpheme compounds, rather than single lexical atoms. As such, Chinese wears its semantic form on its sleeve, in the guise of orthographic choice, and this transparency can be exploited to yield greater semantic insight into concepts. Since the concepts will, by and large, be common to both Chinese and English, these insights can readily be transferred from Chinese to English semantic resources.

[1] School of Computer Science and Informatics, University College Dublin, Ireland, email: {tony.veale, s.chen}@UCD.ie

Chief amongst these insights are the connotational aspects of word meaning. Not every knight is brave, nor every murderer ruthless, yet these are key connotations that must be known by a system if it is to reason about these concepts in a natural, human-like manner. Unfortunately, because connotations are neither definitional nor objective properties of a word or concept, we are unlikely to find them in a lexico-semantic resource like WordNet [5], or even a common-sense knowledge-base like Cyc [3]. Consider the words "violin" and "fiddle": Cruse [1] observes that one cannot imagine a declarative sentence containing one of these words whose truth-conditions would be affected if the other was used in its place. As he notes, violin-playing logically entails fiddle-playing, and vice versa. Notwithstanding this pronouncement, one can nonetheless imagine sentences whose affective meaning, if not their propositional content, is changed when such a substitution is made. The utterance "He is a mere fiddle player" surely loses something in the translation to "He is a mere violin player", most likely because the former communicates a bias founded on the unique connotations of "fiddle" as a musical instrument of the beer-hall rather than of the concert-hall.

In this paper we describe a means of unlocking semantic information from Chinese orthographic forms, so that this information can be transplanted onto English via WordNet [5]. Once transplanted, many of these semantic nuances will reveal new semantic perspectives on concepts common to both languages. In the sense that these perspectives are both novel (to English), and useful (as a source of alternate lexical descriptions), these nuances can be considered truly creative [8]. In this vein, we exploit the most novel of these nuances to generate creative synonyms [6] for existing concepts (such as "ice mountain" for "iceberg" and "fire mountain" for "volcano"), and even to generate creative analogies of the form encountered in the S.A.T. test [7]. In section 2 we describe the necessary resources in more detail, before describing the decomposition and transplant processes in section 3. Potential uses are then described in sections 4 and 5.

2 Lexical Resources

Large-scale lexical resources form the cornerstone of contemporary approaches to Natural Language Understanding (NLU). Of these resources, the most knowledge-rich and labour-intensive to construct are lexical ontologies [2,4,5] - logical structures that attempt to bridge the domain of words and the domain of concepts. Perhaps the most well-known lexical ontology is Princeton WordNet, a broad-coverage electronic thesaurus of English in which word-concepts are organized according to hierarchical (IS-A) and meronymic (part-whole) relationships. An ontology is more than a taxonomy, of course, and WordNet's reliance on hierarchical organization to capture meaning differences marks it as a lightweight ontology, but an

ontology nonetheless. To an extent, more heavyweight ontologies like that of the Cyc [4] project, can also be considered lexical, inasmuch as they explicitly attempt to link the meaning of words to ontological terms. HowNet [2] is a bilingual ontology of Chinese word concepts that has been annotated with the equivalent English translations. It is from HowNet that we obtain the Chinese word-forms that drive this research, while it is WordNet's taxonomic structure that allows us to subject orthographic decompositions of these Chinese word-forms to an English-centric semantic analysis.

As a bilingual English/Chinese lexicon, HowNet allows us to capture the implicit connotational differences that exist between English synonyms by looking to their Chinese translations, where these differences are often explicit. In Chinese, for instance, the concept Lawyer has a connotation of Mastery which is not to be found in WordNet but which is visible in the Chinese word "律师", a concatenation of the characters "律", meaning "law", and "师", meaning "Master". Likewise, the concept Doctor has a connotation of learnedness in Chinese that can be discerned from its Chinese translation, "医生", a conjoining of the ideas characters "医", meaning "medicine", and "生", meaning "pupil". Perceived social status is a nuance not often represented in an explicit lexical semantics. For example, there is nothing intrinsically pejorative about the concept Repairman, yet as a description of a Surgeon the label can appear demeaning. This social gap is visible from a cross-cultural perspective, when we note that the Chinese translation of "repairman", "修理工", is a conjunction of "修理", meaning "to mend", and "工", meaning "worker". It is from the latter character, "工", that repairmen obtains a connotation of the working-, rather than professional-, classes. Social affect can thus be a highly relative and contextual notion, but it can help to quantify the affective difference between otherwise synonymous terms. Consider the words "chef" and "cook": Chinese translates "chef" as "厨师", meaning a "kitchen master", while it translates "cook" as "厨工", meaning a "kitchen worker". Though the word "cook" is not an insult in either English or Chinese, it might well be considered an insult in either language to describe a chef as a cook, just as it might be considered flattery to describe a cook as a chef. Each word concept accentuates a different component of semantic meaning with different dimensions of social meaning.

3 Semantic Form Mirrors Orthographic Form

In a lexical ontology, a compound term - such as "Greek god" or "coffee machine" - represents the yoking of two parts of a concept taxonomy into a single stream. The same can be said even for single-word terms when these words comprise multiple morphemes, though the yoking of domains may be more visible in some languages than in others. For instance, the Chinese word for "espresso" is "浓咖啡", where "浓" can mean either "strong", "rich", "concentrated" or "thick", and "咖啡" means "coffee". In Chinese then, this multi-morpheme word represents a yoking of the HowNet taxonomy of properties with the HowNet taxonomy of entities. By recognizing the nature of this yoke, we can extract explicit *property:value* pairings that can then be grafted onto resources like WordNet.

Chinese character-strings can be decomposed in many different ways, but as one might expect, most dissections do not result in valid semantic analyses. One must be careful to dissect character-strings into meaningful pairs of substrings that describe mutually compatible ideas. As language users, we know that the decomposition of espresso|浓咖啡 into rich|浓 and coffee|咖啡 is a valid one, because richness is a taste setting and coffee, as a kind of beverage, supports the taste property. Unfortunately, this intuition is not supported by HowNet, which neglects to provide a mapping between concepts that express property values and the concepts that can meaningfully hold those values. So we begin by constructing such a high level mapping by hand, by first looking to the HowNet property taxonomy, and for each property type (such as Taste, Courageousness, etc.), we specify the corresponding entity types (such as Foodstuff, Person, etc.) that can exhibit those properties. This high-level mapping allows specific property values (like rich, brave) to be associated with specific entities (like espresso and knights).

To generate well-formed *adjective + noun* decompositions, we employ a hand-coded mapping of 120 property types (such as courageousness) to 50 unique entity types (such as human, animal and beverage); each property type is mapped to an average of 5 entity types. This mapping then allows us to dissect 8290 Chinese noun-forms (denoting 11,219 different senses) in HowNet into a combination of property-value and base-entity, as when knight is dissected to produce a combination of brave + warrior. Interestingly, because most Chinese characters have multiple meanings, most dissections can be given multiple readings, and many of these alternative readings may be taken as valid with respect to the property/entity mapping. Thus, the character 浓 has 8 senses in HowNet, and can denote any of the following property settings: *hue=deep, density=dense, taste=rich, taste=strong, concentration=concentrated, density=thick, intensity=great and intensity=strong*. Since foodstuffs and beverages can support the properties *hue, taste, concentration* and *density*, 6 of these settings are deemed valid with respect to the entry espresso|浓咖啡.

3.1 Evaluation

Chinese nouns in HowNet can be meaningfully decomposed into three different forms: *adjective + noun*, *verb + noun*, and *noun + noun*. In all, we decompose 22,836 multi-character words (denoting 25,343 different senses) can be decomposed into one or more of these categories. The *adjective + noun* pattern accounts for 36% of decompositions, the *verb + noun* pattern for 51%, and *noun + noun* for 45% (clearly then, some words can be decomposed in multiple ways).

One particular decomposition pattern, into gender and base-class, yields a representative "thin slice" of the decomposition process at work. Consider the set of Chinese nouns that yield the property:value pair *sex* (性) = *female* (母/女): mother = female + parent, hen = female + chicken, tigress = female + tiger, virago = female + tiger (a metaphor), pistil = female + stamen, wife = female + person, daughter = female + child, queen = female + monarch, heroine = female + champion, stewardess = female + attendant, actress = female + actor, maidservant = female + servant, bitch = female + dog, mare = female + horse, cow = female + ox, sow = female + hog and lioness = female + lion.

4 Creative Synonym Generation

While reformulations like "strong coffee" have an undeniable explanatory value for describing unknown words, these reformulations can hardly be called creative. Nonetheless, orthographic decomposition yields a whole spectrum of reformulations, some more creative than others. For instance, the orthography of vampire|吸血鬼 permits reformulation as the complex synonym "a ghost (鬼) who sucks (吸) blood (血)". Note that the Chinese character "血" can denote either "blood" as a bodily fluid or "lifeblood" as a animating force, so this reformulation can also be viewed metaphorically as describing any entity that is life-draining.

The real reason that decompositions like “strong coffee” do not seem creative is that their construction is inherently rule-bound, since only combinations that are consistent with our hand-coded mapping of the HowNet property taxonomy are allowed. We need another way of validating creative decompositions, while rejecting nonsense combinations, if creative reformulations are to be deemed valid. Validation must thus be performed against a knowledge source that is both authoritative and flexible: authoritative so that decomposition process is reliable, and flexible so that non-literal acts of creativity are not rejected out of hand. The resource we use here is WordNet [5], which provides both a reasonably rich taxonomy of concepts and a set of textual glosses for these concepts. In effect, we use decomposition to align HowNet with WordNet, following [4], such that a successful alignment indicates a valid decomposition. A decomposition of $\alpha\beta$ into α and β is considered literal w.r.t. WordNet if we can identifying a WordNet sense of $\alpha\beta$ that is a hyponym of some sense of β and whose gloss additionally contains the word α. To allow creative decomposition, we simply weaken the hyponym condition so that $\alpha\beta$ and β merely denote senses that share a common hypernym. Thus, hippo|河马 can validly be decomposed into river|河 + horse|马, dolphin as sea|海 + pig|豚, and zebra|斑马 as striped|斑 + horse|马.

4.1 Evaluation

Strict hypernymic alignment not only stifles creativity, it reduces the applicability of the decomposition process. Using strict alignment between HowNet and WordNet (where some sense of $\alpha\beta$ must be a hyponym of some sense of β), just 3500 multi-character Chinese nouns decompositions are validated. By relaxing this requirement so that $\alpha\beta$ and β merely share a common hypernym at a minimum depth of 3 in the WordNet noun taxonomy, a more expansive set of 11,200 noun decompositions are automatically validated. Naturally, much of the disparity is due to the admission of metaphoric decompositions, though many of these simply verge on the hyperbolic, as when gardener|花匠 is decomposed as flower|花 + artisan|匠. Others exploit the polysemy of individual Chinese logomorphs, as when implication|意味 is decomposed as meaning|意 + flavor|味. Other creative differences are deeply cultural, as in the disparaging use of the concept ghost|鬼 in lie|鬼话 = ghost|鬼 + word|话 and coward|胆小鬼 = timid|胆小 + ghost|鬼.

5 Identifying Metaphors, Analogies and Blends

The decomposition and transplant process reveals many Chinese word forms to be - if not wholly metaphoric - then vaguely analogical in nature. A number of linguistic forces drive this tendency toward the figurative, not least the ancient origins of many Chinese characters and word-forms. Consider that the Chinese concept bone-joint|骨节 is decomposable as skeleton|骨 + knot|节, cervix|宫颈 is decomposable as uterus|宫 + neck|颈 and backbone|骨干 as skeleton|骨 + trunk|干. For similar reasons, electron|电子is decomposable as electricity|电 + seed|子, while robot|机器人 is decomposed as machine|机器 + person|人.

Given the ancient nature of many Chinese character combinations, lexicalized metaphors in Chinese often resemble the *kenning* riddles of old English (in which, for instance, the body is described as a “bone house” and the sky as a “bird house”). Two particularly striking examples are identified by the decomposition processes of section 4: Chinese encodes “breast” (乳房) as a “house|房 of milk|乳”, and “sky” (天宇) as a “celestial|天 house|宇”. Generalizing from these lexicalized metaphors in the context of another language like English should allow a creative system to generate innovative, yet sensible, metaphors of its own.

For the moment, however, analogies can also be derived from orthographic decompositions that are neither analogical or metaphorical, since in general, a lexical analogy can be formed between two decompositions that share a common prefix or head, as in $w_1|\alpha\beta\chi$ = $m|\alpha\beta$ + $h_1|\chi$ and $w_2|\alpha\beta\delta$ = $m|\alpha\beta$ + $h_2|\delta$. The form of the analogy, expressed in the guise of an S.A.T. problem (see [7]) is thus w_1: h_1:: w_2: h_2. For instance, the head element cancer|癌 is common to a number of validated decompositions, suggesting a range of analogies such as *cancroid:skin::adenocarcinoma:gland::seminoma:testis::leukaemia:blood* in each case, the implied relationship is “cancer-type affects body-part”).

Nonetheless, the creativity of each analogy is a function of the insightfulness of the implied relationship, and its ability to draw connections among a heterogeneous set of elements. Many semantic components, like female|母, knowledge|学, source|源 and artisan|匠 are used so frequently as to serve as fruitfully as the pivots of an lexical analogy. However, the most challenging analogies arise from those components that are used in the most diverse contexts. For instance, female|母, is sufficiently metaphoric to be used not just literally, as a sex marker for animate beings, but figuratively, in non-animate concepts such as vowel|母音 = female|母 + sound|音. As such, *enemy:army::antiparticle:particle* is a more creative analogy that the cancer analogies above, since it serves to relate the domains of people and sub-atomic particles.

In addition to metaphor and analogy, conceptual blending is yet another figurative process strongly implicated in Chinese word formation.

6 CONCLUSION

Language, it has been said, is a faculty that makes infinite use of finite resources. Chief among these resources are words, the concepts they denote, and the rules of grammar that allow them to be combined into complex sentences and narratives. Now, while the elements of these resources may be enumerated with some success (think of the utility of dictionaries), fully characterizing these elements is an endeavor that is considerably more open-ended. Remarkable subtlety is demanded of our computational representations if they are to adequately do justice to the chief resources of language - words and their meanings.

We have described a system for mining lexicalized associations, metaphors and analogies from Chinese, a language which wears its conceptual structure relatively openly on its sleeve. In striving for valid decompositions of Chinese lexemes, our approach employs a lexico-semantic touchstone (in the form of Princeton WordNet) that filters out apparently meaningless analyses. But in doing so, it also filters out the most remote, and thus creative, metaphors that Chinese has to offer. For instance, our approach fails to recognize the decomposition tractor|铁牛 = iron|铁 + ox|牛 because the semantic gulf between animals and artifacts makes this decomposition appear spurious. A more knowledge-driven approach to decomposition - such as one that employs specific knowledge of common metaphor families - is needed to resolve this problem. Though still at an early stage of development and inquiry, we believe the current approach sufficiently demonstrates that the structure of one language can be used to reveal a rich array of semantic nuances in another, and that these nuances can be exploited in the generation of creative synonyms, metaphors and analogies.

ACKNOWLEDGEMENTS

We would like to thank Enterprise Ireland's commercialization fund for supporting this research.

REFERENCES

[1] Cruse, A. D., *Lexical Semantics*, Cambridge University Press, London, (1986).

[2] Dong, Z. and Dong, Q., *HowNet and the Computation of Meaning*, World Scientific, Singapore, (2006).

[3] Kim, H., Chen, S. and Veale, T., *Analogical Reasoning with a Synergy of HowNet and WordNet*, In the proceedings of GWC'2006, the 3rd Global WordNet Conference, Cheju, Korea, (January 2006).

[4] Lenat, D. and R. V. Guha., *Building Large Knowledge-Based Systems [CYC]*, Addison Wesley, Reading Massachusetts, (1991).

[5] Miller, G. A., *WordNet: A Lexical Database for English*, Communications of the ACM, 38(11), Amsterdam, (1995).

[6] Veale, T., *The Analogical Thesaurus: An Emerging Application at the Juncture of Lexical Metaphor and Information Retrieval*, Proceedings of IAAI'03, Innovative Applications of Artificial Intelligence, Menlo Park, CA:, AAAI Press, (2003).

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

[8] Boden, M., *Computational models of creativity*, Handbook of Creativity, pp. 351-373, (1999).