Learning to Extract Semantic Content from the Orthographic Structure of Chinese Words

Tony Veale[1] and Shanshan Chen[1]

School of Computer Science and Informatics, University College Dublin, Ireland
{tony.veale, s.chen}@ucd.ie

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. Moreover, the way in which a given language represents complex ideas as aggregations of simpler lexico-semantic components can reveal not just the culture's particular emphases, but can also yield insights about the nature of semantic composition in general.

In this respect, the Chinese written language makes an interesting case in point. Most Chinese words are gestalt terms 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 gestalts, 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. To some extent, lexemes in English are likewise decomposable; for instance, the Latinate origins of words like "reimburse" allows such words to be morphemically decomposed into their underlying ideas, as in "reimburse" = re-im-burse = back-into-bag, or "indisputable" = in-dis-putable = not-apart-think-*adjective*. However, English derives from a multitude of languages and cultures, and the amenability of its lexemes to semantic decomposition is neither as widespread, as noise-free or as transparent as it is for Chinese. Nonetheless, since the concepts in which we are interested will, by and large, be common to both Chinese and English, such decomposition-derived insights can readily be transferred from Chinese to English semantic resources.

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 or 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 [9]. 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 struc-

tures 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 as 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 like 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. Though HowNet's taxonomy of lexical concepts is structurally inferior to WordNet's (it essentially lacks a middle-ontology, instead linking leaf concepts directly to the upper ontology), it compensates for this lack of differentiation by also providing a sparse propositional definition for each concept as follows:

(1) *surgeon*|医生 ≡ {*human*|人: {*doctor*|医治:*agent*={~}}}

This definition can be glossed thus: "a surgeon is a human who acts as an agent of a doctoring activity", where the {~} here serves to indicate the placement of the concept within its associated propositional structure. Because these definitions are lacking in semantic nuance, many non-synonymous terms share the same propositional content in HowNet. However, this lack of precision allows these definitions to serve as complex types, thus serving to unit word-concepts that are functionally similar (often analogously so) rather than strictly identical.

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 may seem 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 Decomposition via Orthographic Analysis

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. However, we believe that such a mapping can be learned automatically, by analyzing the internal structure of Chinese words like espresso|浓咖啡, which yields a mapping from Taste settings to Drink concepts, and warrior|武士, which yields a mapping from Behavioural settings to Person concepts. The bilingual nature of HowNet is essential to this enterprise, since the Chinese word form not only yields the English decomposition but serves to disambiguate this decomposition: “strong” is thus understood in its taste sense (strong|浓 in HowNet) rather than in any physical or mental sense.

3.1 Mapping to Existing Terms

The Chinese character 浓 has 8 different senses in HowNet, and so 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*. Likewise, 士 can denote not just a person, but a scholar, a bachelor and a non-commissioned officer in Chinese. How then do we determine which senses are appropriate for a given gestalt term, or in other words, how do

we determine which of the many possible decompositions are sufficiently compelling to serve as a basis for learning? If we assume that the most acceptable decompositions are those that produce the most natural English collocations, we simply choose those decompositions that yield a lexicalized English compound term. In this regard, both HowNet and WordNet can be used as a source of common English compound terms; HowNet, for instance, contains a lexical entry for "valiant person", allowing us to recognize this as a valid decomposition of "武士". We note that this heuristic is also adept at validating metaphoric decompositions, a class of word associations that would otherwise prove difficult to analyze in purely semantic terms. For instance, the word "战鹰" denotes a fighter plane in Chinese, yet its orthographic form yields the metaphors "war hawk" and "fighting eagle", both of which are stored as lexical items in HowNet.

3.2 Term Recombination

Thusfar, our heuristic for validating English decompositions of Chinese gestalt-words simply exploits the bilingual redundancy of resources like HowNet (or, indeed, the amalgam of HowNet and WordNet we describe in [3]), insofar as many Chinese words can be decomposed into pre-existing English phrases (albeit phrases that may denote different lexical concepts). This approach has obvious limits: for instance, the decomposition "strong coffee" is rejected for 浓咖啡 since neither HowNet nor WordNet specify a meaning for this collocation. However, HowNet does contain entries for "strong tea" (浓茶) and "iced coffee" (冰咖啡), both of which it defines as sub-types of Drink. These examples suggest it should be possible for the decomposition process to learn to recognize novel decompositions, like "strong coffee", as meaningful by creating a semantically-grounded language model from known phrases. That is, from "strong tea" a system can learn that 浓 can denote "strong" in the context of drinks, while from "iced coffee" it can learn that 咖啡 can denote "coffee" in similar contexts. Context is here defined relative to the underlying HowNet propositional definition. Because "strong tea" and "espresso" have isomorphic definitions (i.e., each defines a drink with a particular property; see [10]), it is valid to recombine partial lexicalizations of each to arrive at the alternative lexicalization "strong coffee" for "espresso". Similarly, "rough rice" is validated as a decomposition for "brown rice".

3.3 Creative Synonymy

Chinese orthographic decomposition yields a whole spectrum of insightful reformulations. For instance, the orthography of vampire|吸血鬼 permits reformulation as the complex synonym "a ghost (鬼) who sucks (吸) blood (血)". This particular decomposition is validated by the existence of another term in HowNet, "bloodsucker", that also translates as 吸血鬼. It can be fruitful to decompose Chinese lexemes even when a transparent English translation already exists; for instance, HowNet translates 糙米 as "brown rice" but decomposition reveals a perspective laden with implicit world knowledge, "rough rice".

Likewise, while 草案 translates as both "blueprint" and "preliminary sketch", decomposition suggests another alternative, "rough draft" (which does not exist in HowNet as a lexicalized phrase).

When different lexical-concepts give rise to the same validated decompositions, we have good reason to believe that these lexical-concepts are, if not equivalent, then highly similar. For instance, the decomposition "valiant person" is generated not just for "warrior", but for "knight" and "samurai" as well. While different concepts, this common decomposition suggests that these ideas are structured in the same way, and that bravery is a salient property of each. Likewise, each of "draft", "blueprint", "sketch" and "skeleton" reveals, through decomposition, the salient property of roughness.

4 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 from *cancroid:skin::adenocarcinoma:gland* to *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 [8] is yet another figurative process strongly implicated in Chinese word formation. Of course, conceptual blends are considerably harder to identify than lexical blends (or *portmantea*u words, like "affluenza"), since they represent the integration of different ideas rather than different words. Nonetheless, it is possible to heuristically identify the most common idea blends by examining the orthographic structure of multi-morphemic words. Lexical decomposition reveals that 4227 noun-concepts and 1336 verb-concepts in HowNet can be decomposed into a pair of concepts that share the same direct hypernym, suggesting that these sibling concepts have been blended to create a larger whole. For instance, Chinese defines "burin" as a blend of a "knife" and a "chisel", and "tyrant" as a blend of "king" and "autocrat". The majority of these blends, 66%, are literal, in the sense that the integrated concept shares the same immediate hypernym as its component parts. The remaining 34% are figurative, and contain in their number some intriguing metaphors, such as underling|爪牙 = tooth|爪 + claw|牙 (reminiscent of the English metaphor "tooth and nail"). Most notable, perhaps, is the Chinese formulation of contradiction|矛盾 as spear|矛 + shield|盾, which graphically illustrates a logical abstraction in competing military forces. These combinations demonstrate an emergent quality that is the hallmark of the most sophisticated blends, suggesting that sophisticated conceptual machinery must be brought to bear on their interpretation.

5 Evaluation

The version of HowNet employed in this study contains almost 100,000 lexical entries [2], spanning the categories of noun, verb, adjective and adverb. As noted earlier, HowNet rarely assigns a unique semantic definition to each; rather, each lexical entry shares the same propositional content with an average of 3 other entries [10]. This permits easy generalization across non-identical concepts for the purposes of learning how to validate novel decomposition patterns.

Employing strict validation, only decompositions that correspond to known lexicalized phrases (such as "valiant person" and "war hawk") are considered valid. This strictness limits the number of validated decompositions to just over 5000 phrases. Nonetheless, this set of alternate lexicalizations contains some revealing metaphors. For instance, one sense of the verb "draft" (to compile) yields the metaphoric decomposition "grow grass", since a rough text metaphorically corresponds to a yet-to-be-mown garden.

By learning to partially validate decompositions from multiple exemplars, the scope of validated decompositions is extended considerably, to 24573 lexical

entries (or 25% of the HowNet lexicon). From these decompositions, category mappings can be inferred for 181 different property types in HowNet. For instance, the property Decency is found to characterize a type of expression in 24 different decompositions, a type of show in 21 decompositions, and a type of laughter in 12 decompositions. Similarly, Vulgarity is found to a property of both people and texts, while Accuracy is a property of information, texts and symbols.

The class of Chinese words that combine a gender setting with a base term serves as 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.

6 Conclusions

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 apparently meaningless analyses. But in doing so, it also filters 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 "iron ox" is not a lexicalized metaphor in either HowNet or WordNet. Furthermore, since many metaphoric decompositions of Chinese terms are not semantically anomalous, it is difficult to formulate semantic criteria to recognize metaphors. Rather, a great many 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.

A more knowledge-driven approach to decomposition - such as one that employs specific knowledge of common metaphor families - is thus 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.

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. Veale, T., O Donoghue, D. and Keane, M. T., *Computation and Blending. Cognitive Linguistics*, 11(3/4), pp 253-281, (2000).
9. Boden, M., *Computational models of creativity*, Handbook of Creativity, pp. 351-373, (1999).
10. Veale, T., *Analogy Generation with HowNet*, In the proceedings of IJCAI'2005, the 19th International Joint Conference on Artificial Intelligence, (2005).