When Conset meets Synset: A Preliminary Survey of an Ontological Lexical Resource based on Chinese Characters

Shu-Kai Hsieh
Institute of Linguistics
Academia Sinica
Taipei, Taiwan
shukai@gate.sinica.edu.tw

Chu-Ren Huang
Institute of Linguistics
Academia Sinica
Taipei, Taiwan
churen@gate.sinica.edu.tw

Abstract

This paper describes an on-going project concerning with an ontological lexical resource based on the abundant conceptual information grounded on Chinese characters. The ultimate goal of this project is set to construct a cognitively sound and computationally effective character-grounded machine-understandable resource.

Philosophically, Chinese ideogram has its ontological status, but its applicability to the NLP task has not been expressed explicitly in terms of language resource. We thus propose the first attempt to locate Chinese characters within the context of ontology. Having the primary success in applying it to some NLP tasks, we believe that the construction of this knowledge resource will shed new light on theoretical setting as well as the construction of Chinese lexical semantic resources.

1 Introduction

In the history of western linguistics, writing has long been viewed as a surrogate or substitute for speech, the latter being the primary vehicle for human communication. Such "surrogational model" which neglects the systematicity of writing in its own right has also occupied the predominant views in current computational linguistic studies. This paper is set to provide a quite different perspective along with the Eastern philological tradition of the study of scripts, especially the ideographic one i.e., Chinese characters (Hanzi). We believe that the conceptual knowledge information which has been *grounded* on Chinese characters can be used as a cognitively sound and computationally effective ontological lexical resource in performing some NLP tasks, and it will have contribution to the development of *Semantic Web* as well.

2 Background Issues of Chinese Ideographic Writing

2.1 Ideographic Script and Conceptual Knowledge

From the view of writing system and cognition, human conceptual information has been regarded as being *wired* in ideographic scripts. However, in reviewing the contemporary linguistic literatures concerning with the discussions of the essence of Chinese writing system, we found that the main theoretical dispute lies in the fact that, both structural descriptions and psycholinguistic modeling seem to presume that the notions of *ideography* and *phonography* are mutually exclusive.

To break the theoretical impassé, we take a *pragmatic* position in claiming the tripartite properties of Chinese characters: They are *logographic* (morpho-syllabic) in essence, function *phonologically* at the same time, and can be interpreted *ideographically* and implemented as concept instances by computers.

2.2 Chinese Wordhood

Roughly put, a Chinese character is regarded as an ideographic symbol representing *syllable* and *meaning* of a "morpheme" in spoken Chinese.

But unlike most affixing languages, Chinese has a large class of *morphemes* - which Packard (2000) calls "bound roots" - that possess certain *affixal* properties (namely, they are bound and productive in forming words), but encode **lexical** rather than

grammatical information. These may occur as either the left- or right-hand component of a word. For example, the *morpheme* 輸 (/shu/; "transport") can be used as either the first *morpheme* (e.g., 輸入 (/yùn-rù/; transport-into "import"), or the second *morpheme* (e.g., 運輸 /yùn-shu/; transit-transport "conveyance") of a dissyllabic word, but cannot occur in isolation.

The fuzzy boundary between free and bound morphemes is directly related to the notorious controversial notion of Chinese Wordhood. There are multiple studies showing that to a large extent, (trained or untrained) native speakers of Chinese disagree on what a (free) morpheme/word/compound is.

Such difficulty could be traced back to its historical facts. In modern Mandarin Chinese, there is a strong tendency toward dissyllabic words, while the predominant monosyllabic words in ancient Chinese remain more or less a closed set. But the conceptual knowledge encoded in monosyllabic morphemes still have their influence even on contemporary texts, and thus resulting the difficulties of word-marking decision.

3 Theoretical Setting

Yu et al (1999) reported that a *Morpheme Knowledge Base of Modern Chinese* according to all Chinese characters in GB2312-80 code has been constructed by the institute of Computational Linguistics of Peking University. This Morpheme Knowledge Base has been later integrated into the project called "Grammatical Knowledge Base of Contemporary Chinese".

It is noted that the "morphemes" adopted in this database are monosyllabic "bound morphemes". As for "free morphemes", that is, characters which can be independently used as words, are not included in the Knowledge Base. For example, the *monosyllabic character* 梳 (/shu/,"comb") has (at least) two senses. For the verbal sense ("to comb"), it can be used as a *word*; for the nominal sense ("a comb"), it can only be used in combining with other morphemes. Therefore, only the nominal sense of 梳 is included in the Knowledge Base. However, such *morpheme-based* approach can hardly escape from facing with the difficult decision of free/bound distinction in contemporary Chinese.

3.1 Hanzi/Word Space Model

Based on the consideration mentioned above, in this paper, we will propose a *historical, conventionalized, pre-theoretical* perspective in viewing the lexical and knowledge information within Chinese characters. In Figure 1, (a) illustrates a naive Hanzi space, while (d) shows a linguistic theory-laden result of Hanzi/Word space, where green areas denote to *words*, consisting of 1 to 4 characters. The decision of words (green) and non-words (white) in the space is based on certain perspectives (be it psycholinguistic or computational linguistic). Instead, we take the traditional philological construct of Hanzi into consideration. By analyzing the *conceptual* relations between characters (b) which scatter among diverse lexical resources, we construct an top-level ontology with Hanzi as its instances (c). Rather than $(a) \rightarrow (d)$, which is a predominant approach in contemporary linguistic theoretical construction of Chinese Wordhood, we believe that the proposed approach $(a) \rightarrow (b) \rightarrow (c) \rightarrow (d)$ could not only enclose the implicit conceptual information evolutionarily *encoded* in Chinese characters, but also provide a more clear knowledge scenario for the interaction of characters/words in modern linguistic theoretical setting.

3.2 Conset and Character Ontology

The new model that we propose here is called **HanziNet**. It relies on a novel notion called **conset** and a coarsely grained **upper-level ontology** of characters.

In comparison with synset, which has become a core notion in the construction of Wordnet-like lexical semantic resources, we will argue that there is a crucial difference between Word-based lexical resource and character-based lexical resource, in that they rest with finely-differentiated information contents represented by the nodes of network. A **synset**, or synonym set in WordNet contains a group of words,[1] and each of which is synonymous with the other words in the same synset. In WordNet's design, each synset can be viewed as a *concept* in a taxonomy, While in HanziNet, we are seeking to align Hanzi which share a given putatively primitive meaning extracted from traditional philological resources, so a new term **conset** (concept set) is proposed. A *conset* contains

[1] To put it exactly, it contains a group of lexical units, which can be words or collocations.

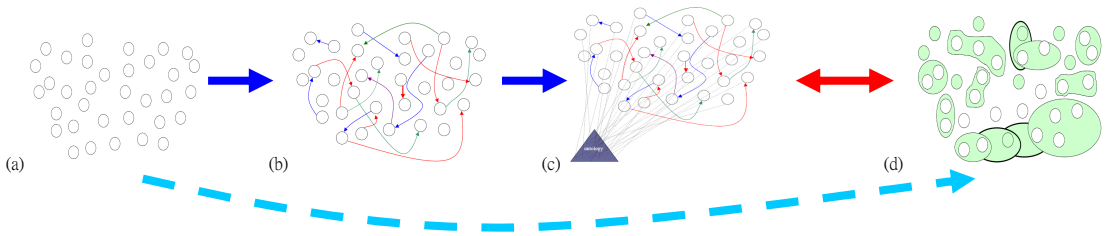

Figure 1: Illustrations of Hanzi/Word Spaces

a group of *Chinese characters similar in concept*, and each of which shares with similar conceptual information with the other characters in the same conset.[2]

The relations between consets constitute a character ontology. Formally, it is a tree-structured conceptual taxonomy in terms of which only two kinds of relations are allowed: the *INSTANCE-OF* (i.e., characters are instances of consets) and *IS-A* relations (i.e., consets are hypernyms/hyponyms to other consets).

Currently, frequently used monosyllabic characters are assigned to *at least* one of 309 consets. Following are some examples:

conset 126 (SUBJECTIVE → EXCITABILITY → ABILITY → ORGANIC FUNCTION)
吸、品、嚐、嚼、嚥、吞、饌、茹、飲,

conset 130 (SUBJECTIVE → EXCITABILITY → ABILITY → SKILLS)
摘、榨、拾、拔、提、攝、選,

conset 133 (SUBJECTIVE → EXCITABILITY → ABILITY → INTELLECT)
牟、謀、考、選、錄、記、聽,

In fact, the core assumption behind the *synset/conset* distinction is non-trivial. In this project, we assume a hypothesis of the *locality* of **Concept Gestalt** and the *context-sensibility* of **Word Sense** concerning with Chinese characters. That is, characters carry two meaning dimensions: on the one hand, they are *lexicalized* concepts; on the other hands, they can be observed linguistically as bound root morphemes and monomorphemic words according to their independent usage in modern Chinese texts.

Figure 2 shows a schematic diagram of our proposed model. In Aitchison's (2003) terms, for the character level, we take an "atomic globule" network viewpoint, where the characters - realized as instances of *core concept Gestalt* - which share similar conceptual information, cluster together. The relationships between these concept Gestalt form a *rooted tree structure*. Characters are thus assigned to the leaves of the tree in terms of an assemblage of bits. For the word level, we take the "cobweb" viewpoint, as *words* -built up from a pool of characters- are connected to each other through lexical semantic relations. In such case, the network does not form a tree structure but a more complex, long-range highly-correlated *random acyclic graphic structure*.

4 Hanzi-grounded Ontological CharacterNet

In light of the previous consideration, this section attempts to further clarify the building blocks of the **HanziNet** system, – a Hanzi-grounded ontological Character Net – with the goal to arrive at a working model which will serve as a framework for ontological knowledge processing. Briefly, HanziNet is consisted of two main parts:

[2] At the time of writing, about 3,600 characters have been finished in their information construction.

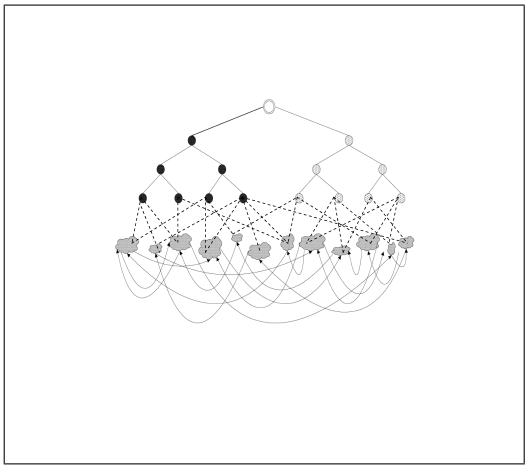

Figure 2: The Schematic Representation of character-triggered tree-like conceptual hierarchy and word-based semantic network

a character-stored machine-readable lexicon and a top-level character ontology.

4.1 Hanzi-grounded Lexicon and Ontology

The current lexicon contains over 5000 characters, and 30,000 derived words in total.[3]

The building of the lexical specification of the entries in HanziNet includes various aspects of Hanzi:

1. *Conset(s)*: The conceptual code is the core part of the MRD lexicon in HanziNet. Concepts in HanziNet are indicated by means of a label (conset name) with a code form. In order to increase the efficiency, an ideal strategy is to adopt the Huffmann-coding-like method, by encoding the conceptual structure of Hanzi as a pattern of bits set within a bit string.[4] The *coding* thus refers to the assignment of code sequences to an character. The sequence of edges from the root to any character yields the code for that character, and the number of bits varies from one character to another. Currently, for each conset (309 in total) there are 12 characters assigned on the average; for each character, it is assigned to 2-3 consets on the average.[5]
2. *Character Semantic Head (CSH) and Character Semantic Modifier (CSM) division.*[6]
3. Shallow parts of speech (mainly Nominal(N) and Verbal(V) tags)
4. Gloss of *prototypical meaning*
5. List of *combined words with statistics calculated from corpus*, and
6. Further aspects such as *character types and cognates*: According to ancient study, characters can be compartmentalized into six groups based on the six classical principles of character construction. Character type here means which group the character belongs to. And the term *cognate* here is defined as characters that share the same *CSH* or *CSM*. Figure 3 shows a snapshot of this lexicon.

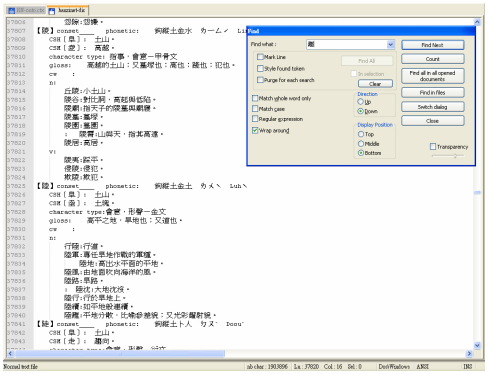

Figure 3: The character-stored lexicon: a snapshot

The second core component of the proposed resource is a set of hierarchically related *Top Concepts* called *Top-level Ontology* (or Upper ontology). This is similar to EuroWordnet 1.2, which is

[3]Since this lexicon aims at establishing an knowledge resource for modern Chinese NLP, characters and words are mostly extracted from the Academia Sinica Balanced Corpus of Modern Chinese (http://www.sinica.edu.tw/SinicaCorpus/), those characters and words which have probably only appeared in classical literary works, (considered *ghost words* in the lexicography), will be discarded.

[4]This is inspired by Chu (1999)'s works.

[5]The disputing point here is that, if some of the monosyllabic morphemes are taken as *words*, they should be very ambiguous in the daily linguistic context, at least more ambiguous than the dissyllabic words. However, as we argued previously, HanziNet takes a different perspective in locating theoretical roles of Hanzi.

[6]This distinction is made based on the glyphographical consideration, which has been a crucial topic in the studies of traditional Chinese scriptology. Due to the limited space, this will not be discussed here.

also enriched with the *Top Ontology* and the set of *Base Concepts* (Vossen 1998).

As mentioned, a tentative set of 309 *conset*, a kind of ontological categories in contrast with synset has been proposed [7], and over 5000 characters have been used as instances in populating the character ontology.

Methodologically, following the basic line of *OntoClear* approach (Guarino and Welty (2002)), we use *simple monotonic inheritance* in our ontology design, which means that each node inherits properties only from a single ancestor, and the inherited value cannot be overwritten at any point of the ontology. The decision to keep the relations to one single parent was made in order to guarantee that the structure would be able to grow indefinitely and still be manageable, i.e. that the transitive quality of the relations between the nodes would not degenerate with size. Figure 4 shows a snapshot of the character ontology.

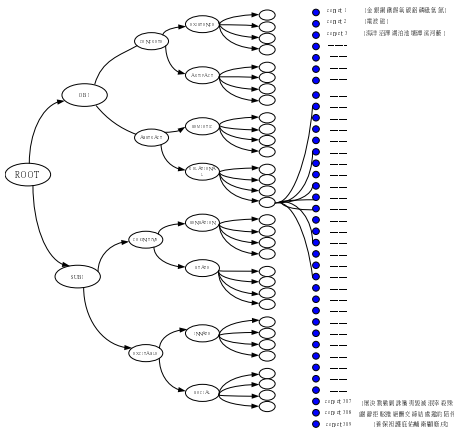

Figure 4: The character ontology: a snapshot

4.2 Characters in a Small World

In addition, an experiment concerning the *character network* that was based on the meaning aspects of characters, was performed from a statistical point of view. It was found that this character network, like many other linguistic semantic networks (such as WordNet), exhibits a *small-world* property (Watt 1998), characterized by sparse connectivity, small average shortest paths between characters, and strong local clustering. Moreover, due to its dynamic property, it appears to exhibit an asymptotic *scale-free* (Barabasi 1999) feature with the connectivity of power laws distribution, which is found in many other network systems as well.

[7] It would be interesting to compare consets with the basic 400 nodes in the upper region proposed by Hovy(2005).

Table 1: Statistical characteristics of the character network: $\mathcal{N}$ is the total number of nodes(characters), $\overline{k}$ is the average number of links per node, $\mathcal{C}$ is the clustering coefficient, and $\overline{L}$ is the average shortest-path length, and L_{max} is the maximum length of the shortest path between a pair of characters in the network.

	N	$\overline{k}$	$\mathcal{C}$	$\overline{\mathcal{L}}$
Actual configuration	6493	350	0.64	2.0
Random configuration	6493	350	0.06	1.5

Our first result is that our proposed conceptual network is highly clustered and at the same time and has a very small length, i.e., it is a small world model in the *static* aspect. Specifically, $\mathcal{L} \gtrsim \mathcal{L}_{random}$ but $\mathcal{C} \gg \mathcal{C}_{random}$. Results for the network of characters, and a comparison with a corresponding random network with the same parameters are shown in Table 1. $\mathcal{N}$ is the total number of nodes (characters), $\overline{k}$ is the average number of links per node, $\mathcal{C}$ is the clustering coefficient, and $\mathcal{L}$ is the average shortest path.

4.3 HanziNet in the Global Wordnet Grid

In order to promote a semantic and ontological interoperability, we have aligned conset with the 164 Base Concepts, a shared set of concepts from EWN in terms of Wordnet synsets and SUMO definitions, which has been currently proposed in the international collaborative platform of *Global Wordnet Grid*.

5 Applications and Future Development

5.1 Sense Prediction and Disambiguation

Based on the initial version of the proposed resources, Hsieh (2005b) has proposed a semantic class prediction model which aims to gain the possible semantic classes of unknown two-characters words. The results obtained shows that, with this knowledge resource, the system can achieve fairly high level of performance. Meaning relevant NLP Tasks such as Word Sense Disambiguation are also in preparation.

5.2 Interfacing Hantology, HanziNet and Chinese Wordnet

Interfacing ontologies and lexical resources has been a research topic in the coming age of semantic web. In the case of Chinese, three existing lexical resources (意符 Radicals::Hantology (Chou and Huang (2005))- 字 Characters::HanziNet - 詞 Words::Chinese Wordnet) constitutes an integrated 3-level knowledge scenario which would provide important insights into the problems of understanding the complexities and its interaction with Chinese natural language.

6 Conclusion

In conclusion, the goal of this research is set to survey the unique characteristics of Chinese Ideographs.

Though it has been well understood and agreed upon in cognitive linguistics that concepts can be represented in many ways, using various constructions at different syntactical levels, conceptual representation at the script level has been unfortunately both undervalued and under-represented in computational linguistics. Therefore, the Hanzi-driven conceptual approach in this thesis might require that we consider the Chinese writing system from a perspective that is not normally found in canonical treatments of writing systems in contemporary linguistics.

Against the deep-seated tradition in contemporary Chinese linguistics, which views the use of Chinese characters in scientific theories as a manifestation of mathematical immaturity and interpretational subjectivity, we propose the first lexical knowledge resource based on Chinese characters in the field of linguistic as well as in the NLP.

It is noted that HanziNet, as a general knowledge resource, should not claim to be a sufficient knowledge resource in and of itself, but instead seek to provide a groundwork for the incremental integration of other knowledge resources for language processing tasks. In order to augment HanziNet, additional information will needed to be incorporated and mapped into HanziNet. This leads us to several avenues of future research.

Acknowledgements

The authors would like to thank the anonymous referees for constructive comments. Thanks also go to the institute of linguistics of Academia Sinica for their kindly data support.

References

Aitchison, Jean. 2003. Words in the mind: an introduction to the mental lexicon. Blackwell publishing.

Barabasi, Albert-Laszlo and Reka Albert. 1999. Emergence of scaling in random networks. *Science*, 286:509-512.

Chou, Ya-Min and Chu-Ren Huang. 2005. Hantology: An ontology based on conventionalized conceptualization. *OntoLex Workshop*, Korea.

Chu, Bong-Foo. 1999-. http://www.cbflabs.com

Guarino, Nicola and Chris Welty. 2002. Evaluating ontological decisions with OntoClean. In: *Communications of the ACM*. 45(2):61-65

Hovy, E.H. 2005. Methodologies for the Reliable Construction of Ontological Knowledge. In : F. Dau, M.-L. Mugnier, and G. Stumme (eds), *Conceptual Structures: Common Semantics for Sharing Knowledge.* Proceedings of the 13th Annual International Conference on Conceptual Structures (ICCS 2005). Kassel, Germany.

Hsieh, Shu-Kai. 2005(a). HanziNet: An enriched conceptual network of Chinese characters. *The 5rd workshop on Chinese lexical semantics*, China: Xiamen.

Hsieh, Shu-Kai. 2005(b). Word Meaning Inducing via Character Ontology. *IJINLP, SIGHAN Workshop*, Jijeu Island, South Korea.

Packard, J. L. 2000. *The morphology of Chinese*. Cambridge, UK: Cambridge University Press.

Steyvers, M. and Tenenbaum, J.B. 2002 The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*.

Watts, D. J. and Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature* 393:440-42.

Yu, Shiwen, Zhu Xuefeng and Li Feng. 1999. The development and application of modern Chinese morpheme knowledge base.[in Chinese]. In: 世界漢語教學, No.2. pp38-45.