Word Meaning Inducing via Character Ontology: A Survey on the Semantic Prediction of Chinese Two-Character Words

Shu-Kai Hsieh
Seminar für Sprachwissenschaft
Abt. Computerlinguistik
Universität Tübingen
72074, Germany
kai@hanzinet.org

Abstract

This paper presents a semantic class prediction model of Chinese two-character compound words based on a character ontology, which is set to be a feasible conceptual knowledge resource grounded in Chinese characters. The experiment we conduct yields satisfactory results which turn out to be that the task of semantic prediction of two-character words could be greatly facilitated using Chinese characters as a knowledge resource.

1 Introduction

This paper describes the theoretical consideration concerning with the interaction of ontology and morpho-semantics, and an NLP experiment is performed to do semantic class prediction of unknown two-character words based on the ontological and lexical knowledge of Chinese morphemic components of words (i.e., characters). The task that the semantic predictor (or classifier) performs is to automatically assign the (predefined) semantic thesaurus classes to the unknown two-character words of Chinese.

Among these types of unknown words, Chen and Chen (2000) pointed out that *compound words* constitute the most productive type of unknown words in Chinese texts. However, the caveat at this point should be carefully formulated, due to the fact that there are no unequivocal opinions concerning with some basic theoretical settings in Chinese morphology. The notion of *word*, *morpheme* and *compounding* are not exactly in accord with the definition common within the theoretical setting of Western morphology. To avoid unnecessary misunderstanding, the pre-theoretical term *two-character* words will be mostly used instead of *compound* words in this paper.

2 Word Meaning Inducing via Character Meaning

2.1 Morpho-Semantic Description

As known, "bound roots" are the largest classes of *morpheme* types in Chinese morphology, and they are very productive and represent lexical rather than grammatical information (Packard 2000). This morphological phenomena leads many Chinese linguists to view the word components (i.e., characters) as building blocks in the *semantic composition* process of dis- or multisyllabic words. In many empirical studies (Tseng and Chen (2002); Tseng (2003); Lua (1993); Chen (2004)), this view has been confirmed repeatedly.

In the semantic studies of Chinese word formation, many descriptive and cognitive semantic approaches have been proposed, such as argument structure analysis (Chang 1998) and the frame-based semantic analysis (Chu 2004). However, among these qualitative explanation theoretical models, problems often appear in the *lack of predictability* on the one end of spectrum, or *overgeneration* on the other.[1] Empirical data have

[1] For example, in applying Lieber's (1992) analysis of argument structure and theta-grid in Chinese V-V compounds, Chang (1998) found some examples which may satisfy the semantic and syntactic constraints, but they may not be ac-

also shown that in many cases, – e.g., the abundance of phrasal lexical units in any natural language, – the principle of *compositionality* in a strict sense, that is, “the meaning of a complex expression can be *fully* derivable from the meanings of its component parts, and from the schemas which sanction their combination”(Taylor 2002), which is taken to be a fundamental proposition in some of *morpho-semantically motivated* analysis, is highly questionable.

This has given to the consideration of the embeddedness of linguistic meanings within broader *conceptual structures*. In what follows, we will argue that an *ontology-based* approach would provide an interesting and efficient prospective toward the *character-triggered* morpho-semantic analysis of Chinese words.

2.2 Conceptual Aggregate in *Compounding*: A Shift Toward Character Ontology

In prior studies, it is widely presumed that the category (be it syntactical or semantic) of a word, is somehow strongly associated with that of its composing characters. The *semantic compositionality* underlying two-character words appears in different terms in the literature.[2]

Word semantic similarity calculation techniques have been commonly used to retrieve the similar compositional patterns based on semantic taxonomic thesaurus. However, one weak point in these studies is that they are unable to separate *conceptual* and *semantic* levels. Problem raises when words in question are *conceptually* correlated are not necessarily *semantically* correlated, viz, they might or might not be physically close in the CILIN thesaurus (Mei et al 1998). On closer observations, we found that most synonymic words (i.e., with the same CILIN semantic class) have characters which carry similar conceptual information. This could be best illustrated by examples. Table 1 shows the *conceptual distribution* of the *modifiers* of an example of VV compound by presuming the second character 取 as a *head*. The first column is the semantic class of CILIN (middle level), the second column lists the instances with lower level classification number, and the third column lists their conceptual types adopted from a character ontology we will discuss later. As we can see, though there are 12 resulting semantic classes for the * 取 compounds, the modifier components of these compounds involve only 4 concept types as follows:

11000 (SUBJECTIVE → EXCITABILITY → ABILITY → ORGANIC FUNCTION) 吸、攝,

11010 (SUBJECTIVE → EXCITABILITY → ABILITY → SKILLS) 摘、榨、拾、拔、提、攝、選,

11011 (SUBJECTIVE → EXCITABILITY → ABILITY → INTELLECT) 牟、謀、考、選、錄、記、聽,

11110 (SUBJECTIVE → EXCITABILITY → SOCIAL EXPERIENCE → DEAL WITH THINGS) 收、獲、領、換、奪、竊、詐、獵、掠、爭、剽、襲、攻、進

We defined these patterns as *conceptual aggregate pattern* in compounding. Unlike statistical measure of the co-occurrence restrictions or association strength, a *concept aggregate pattern* provides a more knowledge-rich scenario to represent a specific manner in which concepts are aggregated in the ontological background, and how they affect the compounding words. We will propose that the semantic class prediction of Chinese two-character words could be improved by making use of their *conceptual aggregate pattern* of head/modifier component.

3 Semantic Prediction of Unknown Two-Character Words

The practical task intended to be experimented here involves the automatic classification of Chinese two-character words into a predetermined number of semantic classes. Difficulties encountered in previous researches could be summarized as follows:

First, many models (Chen and Chen 1998;2000) cannot deal with the issue of “incompleteness” of characters in the lexicon, for these models depend heavily on CILIN, a Chinese Thesaurus containing only about 4,133 monosyllabic morphemic components (characters). As a result, if unknown words contain characters that are not listed in CILIN, then the prediction task cannot be performed automatically. Second, the ambiguity of characters is often shunned by

---

ceptable to native speakers.

[2]Using statistical techniques, Lua (1993) found out that each Chinese two-character word is a result of 16 types of *semantic transformation patterns*, which are extracted from the meanings of its constituent characters. In Chen (2004), the combination pattern is referred to as *compounding semantic template*.

SC	VV compounds	Concept types of modifier component
Ee	37 進取	11110
Fa	05 榨取 08 摘取 15 拾取	11010
Fc	05 聽取	11011
Gb	07 記取	11011
Ha	06 奪取	11110
Hb	08 襲取 12 攻取 12 奪取 12 襲取	11110
Hc	07 收取 23 拔取 25 錄取	{11110;11011}
Hi	27 領取 27 提取	{11010;11110}
Hj	25 選取 25 摘取	{11010;11110}
Hn	03 掠取 10 剽取 12 榨取	11110
If	09 考取	11011
Je	12 爭取 12 詐取 12 吸取 12 攝取 12 竊取 12 牟取 12 謀取 12 掠取 12 獵取 12 截取 12 獲取 12 換取 12 奪取	{11000;11110;11011;11110}

Table 1: Conceptual aggregate patterns in two-character VV (compound) words: An example of *取

manual pre-selection of character meaning in the training step, which causes great difficulty for an automatic work. Third, it has long been assumed (Lua 1997; Chen and Chen 2000) that the overwhelming majority of Chinese compounds are more or less *endocentric*, where the compounds denote a hyponym of the *head* component in the compound. E.g, 電郵 ("electric-mail"; e-mail) is a kind of mail. So the process of identifying semantic class of a *compound* boils down to find and to determine the semantic class of its *head* morpheme. However, there is also an amount of *exocentric* and appositional compounds[3] where no straightforward criteria can be made to determine the head component. For example, in a case of VV compound 訓斥 ("denounce-scold", drop-on), it is difficult (and subjective) to say which character is the head that can assign a semantic class to the compound.

To solve above-mentioned problems, Chen (2004) proposed a non head-oriented character-sense association model to retrieve the *latent* senses of characters and the *latent* synonymous compounds among characters by measuring similarity of semantic template in compounding by using a MRD. However, as the author remarked in the final discussion of classification errors, the performance of this model relies much on the productivity of compounding semantic templates of the target compounds. To correctly predict the semantic category of a compound with an unproductive semantic template is no doubt very difficult due to a sparse existence of the template-similar compounds. In addition, the statistical measure of sense association does not tell us any more about the constraints and knowledge of conceptual combination.

In the following, we will propose that a knowledge resource at the morpheme (character) level could be a straightforward remedy to these problems. By treating characters as instances of conceptual primitives, a character ontology thereof might provide an interpretation of *conceptual grounding* of word senses. At a coarse grain, the character ontological model does have advantages in efficiently defining the *conceptual space* within which character-grounded concept primitives and their relations, are *implicitly* located.

4 A Proposed Character Ontology-based Approach

In carrying out the semantic prediction task, we presume the *context-freeness* hypothesis, i.e., without resorting to any contextual information. The consideration is taken based on the observation that native speaker seems to reconstruct their new conceptual structure *locally* in the processing of unknown compound words. On the other hand, it has the advantage especially for those unknown words that occur only once and hence have limited context.

In general, the approach proposed here differs in some ways from previous research based on the following presuppositions:

[3] Lua reports a result of 14.14% (Z3 type).

4.1 Character Ontology as a Knowledge Resource

The new model that we will present below will rely on a coarsely grained upper-level ontology of characters.[4] This character ontology is a tree-structured conceptual taxonomy in terms of which only two kinds of relations are allowed: the *INSTANCE-OF* (i.e., certain characters are instances of certain concept types) and *IS-A* relations (i.e., certain concept type is a kind of certain concept type).

In the character ontology, monosyllabic characters [5] are assigned to *at least* [6] one of 309 consets (concept set), a new term which is defined as a type of concept sharing a given putatively primitive meaning. For instance, 説 (speak), 道 (chatter), 曰 (say), 云 (say), 告 (tell), 訴 (inform), 講 (explain), 敘 (narrate), 謂 (be called), 述 (state), these characters are assigned to the same conset.

Following the basic line of *OntoClear* methodology (Guarino and Welty (2002)), we use *simple monotonic inheritance*, 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 1 shows a snapshot of the character ontology.

4.2 Character-triggered Latent Near-synonyms

The rationale behind this approach is that similar conceptual primitives - in terms of characters - probably participate in similar context or have similar *meaning-inducing* functions. This can be rephrased as the following presumptions: (1). Near-synonymic words often overlap in senses, i.e., they have same or close semantic classes. (2). Words with characters which share similar conceptual information tend to form a *latent* cluster

[4] At the time of writing, about 5,600 characters have been finished in their information construction. Please refer to [4]

[5] In fact, in addition to monosyllabic morpheme, it also contains a few dissyllabic morphemes, and *borrowed polysyllabic morphemes.*

[6] This is due to the homograph.

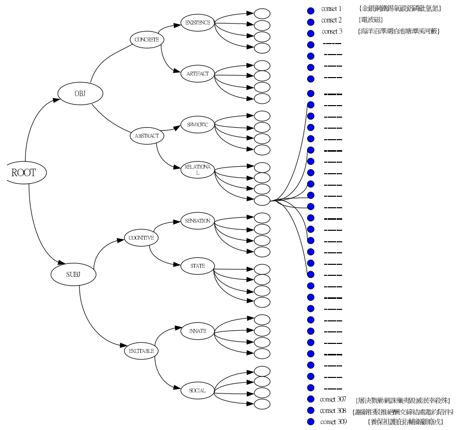

Figure 1: The character ontology: a snapshot

of synonyms. (2). These similar conceptual information can be formalized as *conceptual aggregate patterns* extracted from a character ontology. (3). Identifying such *conceptual aggregate patterns* might thus greatly benefit the automatically acquired near-synonyms, which give a set of good candidates in predicting the semantic class of previously unknown ones.

The proposed semantic classification system retrieves at first a set of near-synonym candidates using *conceptual aggregation patterns*. Considerations from the view of lexicography can winnow the *overgenerated* candidates, that is, a final decision of a list of near-synonym candidates is formed on the basis of the CILIN's verdict as to what latent near-synonyms are. Thus the semantic class of the target unknown two-character words will be assigned with the semantic class of the top-ranked near-synonym calculated by the similarity measurement between them. This method has advantage of avoiding the snag of apparent multiplicity of semantic usages (ambiguity) of a character.

Take for an example. Suppose that the semantic class of a two-character word 保護 (protect; Hi37) is unknown. By presuming the leftmost character 保 the *head* of the word, and the rightmost character 護 as the *modifier* of the word, we first identify the conset which the modifier 護 belongs to. Other instances in this conset are 保, 衵, 戍, 衛, 庇, 佑, 顧, 輔, 佐, 守, 養, etc. So we can retrieve a set of possible *near-synonym* candidates by substitution, namely, NS_1: {保保, 保衵, 保戍, 保衛, 保庇, 保佑, 保顧, 保輔, 保佐, 保守, 保養}; in the same way, by presuming 護 as the *head*, we have a second set of possible *near-synonym* candidates, NS_2: {護護, 衵護, 戍護, 衛護, 庇護, 佑護, 顧護, 輔護, 佐護, 守護, 養護}[7]. Aligned with CILIN, those candidates which are also listed in the CILIN are adopted as the final two list of the near-synonym candidates for the unknown word 保護: NS'_1: {衵護(Hi41), 衛護(Hb04;Hi37), 庇護(Hi47), 守護(Hi37),養護(Hd01)}, and NS'_2: {保佑(Hl33),保養(Hj33), 保守(Ee39)}.

4.3 Semantic Similarity Measure of Unknown Word and its Near-Synonyms

Given two sets of character-triggered near-synonyms candidates, the next step is to calculate the semantic similarity between the unknown word (UW) and these near-synonyms candidates.

CILIN Thesaurus is a tree-structured taxonomic semantic structure of Chinese words, which can be seen as a special case of semantic network. To calculate semantic similarity between nodes in the network can thus make use of the structural information represented in the network.

Following this information content-based model, in measuring the semantic similarity between unknown word and its candidate near-synonymic words, we use a measure metric modelled on those of Chen and Chen (2000), which is a simplification of the Resnik algorithm by assuming that the occurrence probability of each leaf node is equal. Given two sets (NS'_1, NS'_2) of candidate near synonyms, each with m and n near synonyms respectively, the similarity is calculated as in equation (1) and (2), where sc_{uwc1} and sc_{uwc2} are the semantic class(es) of the first and second morphemic component (i.e., character) of a given unknown word, respectively. sc_i and sc_j are the semantic classes of the first and second morphemic components on the list of candidate near-synonyms NS'_1 and NS'_2. f is the frequency of the semantic classes, and the denominator is the total value of numerator for the purpose of normalization. β and $1-\beta$ are the weights which will be discussed later. The *Information Load* (IL) of a semantic class sc is defined in Chen and Chen (2004):

$$IL(sc) = Entropy(system) - Entropy(sc) \quad (3)$$

$$\begin{aligned} &\simeq (-\frac{1}{q}\sum \log_2 \frac{1}{q}) - (-\frac{1}{p}\sum \log_2 \frac{1}{p}) \\ &= \log_2 q - \log_2 p \\ &= -\log_2(\frac{p}{q}), \end{aligned}$$

if there is q the number of the minimal semantic classes in the system,[8] p is the number of the semantic classes subordinate sc.

4.4 Circumventing "Head-oriented" Presupposition

As remarked in Chen (2004), the previous research concerning the automatic semantic classification of Chinese compounds (Lua 1997; Chen and Chen 2000) presupposes the *endocentric* feature of compounds. That is, by supposing that compounds are composed of a *head* and a *modifier*, determining the semantic category of the *target* therefore boils down to determine the semantic category of the *head* compound.

In order to circumventing the strict "head-determination" presumption, which might suffer problems in some borderline cases of V-V compounds, the weight value (β and $1-\beta$) is proposed. The idea of weighting comes from the discussion of *morphological productivity* in Baayen (2001). We presume that, within a given two-character words, the more productive, that is, the more numbers of characters a character can combine with, the more possible it is a *head*, and the more weight should be given to it. The weight is defined as $\beta = \frac{C(n,1)}{N}$, viz, the number of candidate *morphemic components* divided by the total number of N. For instance, in the above-mentioned example, NS_1 should gain more weights than NS_2, for 護 can combine with more characters (5 near-synonyms candidates) in

[7] Note that in this case, 保 and 護 are happened to be in the same conset.

[8] In CILIN, $q = 3915$.

$$sim_{\mu}(UW, NS'_1) = \arg_{i=1,m}^{max} \frac{IL(LCS(sc_{uwc1}, sc_i)) * f_i}{\sum_{i=1}^{m} IL(LCS(sc_{uwc1}, sc_i)) * f_i}(\beta) \quad (1)$$

$$sim_{\nu}(UW, NS'_2) = \arg_{j=1,n}^{max} \frac{IL(LCS(sc_{uwc2}, sc_j)) * f_j}{\sum_{j=1}^{n} IL(LCS(sc_{uwc2}, sc_j)) * f_j}(1-\beta) \quad (2)$$

NS_1 than 保 does in NS_2 (3 near-synonyms candidates). In this case, $\beta = \frac{5}{8} = 0.625$. It is noted that the weight assignment should be character and position independent.

4.5 Experimental Settings

4.5.1 Resources

The following resources are used in the experiments: (1)Sinica Corpus[9], (2) CILIN Thesaurus (Mei et al 1998) and (3) a Chinese character upper-level ontology.[10] (1) is a well known balanced Corpus for modern Chinese used in Taiwan. (2) CILIN Thesaurus is a Chinese Thesaurus widely accepted as a semantic categorization standard of Chinese word in Chinese NLP. In CILIN, a collection of about 52,206 Chinese words are grouped in a Roget's Thesaurus-like structure based on *categories* within which there are several 3 levels of finer clustering (12 major, 95 minor and 1428 minor semantic classes).(3) is an on-going project of Hanzi-grounded Ontology and Lexicon as introduced.

4.5.2 Data

We conducted an open test experiment, which meant that the training data was different from the testing data. 800 two-character words in CILIN were chosen at random to serve as test data, and all the words in the test set were assumed to be unknown. The distribution of the grammatical categories of these data is: NN (200, 25%), VN (100, 12.5%) and VV (500, 62.5%).

4.5.3 Baseline

The baseline method assigns the semantic class of the randomly picked head component to the semantic class of the unknown word in question. It is noted that most of the morphemic components (characters) are ambiguous, in such cases, semantic class is chosen at random as well.

[9]http://www.sinica.edu.tw/SinicaCorpus/
[10]http://www.hanzinet.org/HanziOnto/

Compound types	Baseline	Our algorithm
V-V	12.20%	42.00%
V-N	14.00%	37.00%
N-N	11.00%	72.50%

Table 2: Accuracy in the test set (level 3)

4.5.4 Outline of the Algorithm

Briefly, the strategy to predict the semantic class of a unknown two-character word is, to measure the semantic similarity of unknown words and their candidate near-synonyms which are retrieved based on the character ontology. For any unknown word UW, which is the character sequence of C_1C_2, the $RANK(sim_{\mu}(\beta), sim_{\nu}(1-\beta))$ is computed. The semantic category sc of the candidate synonym which has the value of $MAX(sim_{\mu}(\beta), sim_{\nu}(1-\beta))$, will be the top-ranked guess for the target unknown word.

4.6 Results and Error Analysis

The correctly predicted semantic class is the sematic class listed in CILIN. In the case of ambiguity, when the unknown word in question belongs to more than one semantic classes, any one of the classes of an ambiguous word is considered correct in the evaluation.

The SC prediction algorithm was performed on the test data for outside test in level-3 classification. The resulting accuracy is shown in Table 2. For the purpose of comparison, Table 3 also shows the more shallow semantic classification (the 2nd level in CILIN).

Generally, without contextual information, the classifier is able to predict the meaning of a Chinese two-character words with satisfactory accu-

Compound types	Baseline	Our algorithm
V-V	13.20%	46.20%
V-N	16.00%	42.00%
N-N	12.50%	76.50%

Table 3: Accuracy in the test set (level 2)

racy against the baseline. A further examination of the bad cases indicates that error can be grouped into the following sources:

- Words with no semantic transparency:
  Like "proper names", these types have no semantic transparency property, i.e., the word meanings can not be derived from their morphemic components. Loan words such as 沙發 (/shāfā/; "sofa") are typical examples.

- Words with weak semantic transparency:
  These can be further classified into four types:
  - Appositional compounds: words whose two characters stand in a coordinate relationship, e.g. 東西 ("east-west", thing).
  - Lexicalized idiomatic usage: For such usage, each word is an indivisible construct and each has its meaning which can hardly be computed by adding up the separate meaning of the components of the word. The sources of these idiomatic words might lie in the *etymological past* and are at best meaningless to the modern native speaker. e.g, 薪水 ("salary-water", salary).
  - Metaphorical usage: the meaning of such words are therefore different from the literal meaning. Some testing data is not semantically transparent due to their metaphorical uses, For instance, 喉舌 (Aj) is assigned to the 喉頭 (Bk).

- Derived words:
  Such as 進去 (enter). These could be filter out using syntactical information.

- The quality and coverage of CILIN and character ontology:
  Since our SC system's test and training data are gleaned from CILIN and the character ontology, the quality and coverage play a crucial role. For example, for the unknown compound word 慅慅 (/sāo-sāo/; "be in tumult"), there not even an example which has 慅 as the first character or as the second character. the same problem such as falling short on coverage and data sparseness goes to the character ontology, too. For instance, there are some dissyllabic morphemes which are not listed in ontology, such as 覬覦 (/jìyú/;"covet").

Compound types	Our model	Current best model
V-V	42.00%	39.80% (Chen 2004)
N-N	72.50%	81.00% (Chen and Chen 2000)

Table 4: Level-3 performance in the outside test: a comparison

4.7 Evaluation

So far as we know, no evaluation in the previous works was done. This might be due to many reasons: (1) the different scale of experiment (how many words are in the test data?), (2) the selection of syntactic category (VV, VN or NN?) of morphemic components, and (3) the number of morphemic components involved (two or three-character words?).. etc. Hence it is difficult to compare our results to other models. Among the current similar works, Table 4 shows that our system outperforms Chen(2004) in VV compounds, and approximates the Chen and Chen(2000) in NN compounds.

5 Conclusion

In this paper, we propose a system that aims to gain the possible semantic classes of unknown words via similarity computation based on character ontology and CILIN thesaurus. In general, we approach the task in a hybrid way that combines the strengths of ontology-based and example-based model to achieve at better result for this task.

The scheme we use for automatic semantic class prediction takes advantage of the presumptions that the conceptual information *wired* in Chinese characters can help retrieve the near-synonyms, and the near-synonyms constitute a key indicator for the semantic class guess of unknown words in question.

The results obtained show that, our SC prediction algorithm can achieve fairly high level of performance. While the work presented here in still in progress, a first attempt to analyze a test set of 800 examples has already shown a 43.60% correctness for VV compounds, 41.00% for VN compounds, and 74.50% for NN compounds at the level-3 of CILIN. If shallow semantics is taken into consideration, the results are even better.

Working in this framework, however, one point as suggested by other similar approach is that, human language processing is not limited to an abstract ontology alone (Hong et al. 2004). In practical applications, ontologies are seldom used as the only knowledge resources. For those unknown words with very weak semantic transparency, it would be interesting to show that an ontology-based system can be greatly boosted when other information sources such as metaphor and etymological information integrated. Future work is aimed at improving this accuracy by adding other linguistic knowledge sources and extending the technique to WSD (Word Sense Disambiguation).

Acknowledgements

I would like to thank Erhard Hinrichs and Lothar Lemnitzer for their useful discussions. I also thank the anonymous referees for constructive comments. Thanks also go to the institute of linguistics of Academia Sinica for their kindly data support.

References

Baayen, Harald. (2001). Word frequency distributions. Kluwer Academic Publishers.

Chen, Keh-Jiann and Chao-Jan Chen. (2000). Automatic semantic classification for Chinese unknown compound nouns. COLING 2000, Saarbrücken, Germany.

Chen, Chao-Ren. (2004). Character-Sense association and compounding template similarity: Automatic semantic classification of Chinese compounds. *The 3rd SIGHAN Workshop*.

Chu, Yan. (2004). *Semantic word formation of Chinese compound words*. Peking University Press.

HanziNet Project: `http://www.hanzinet.org`.

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

Hong, Li and Huang (2004). Ontology-based Prediction of Compound Relations: A study based on SUMO. *PACLIC 18.*

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

Lin, Dekang. (1998). A information-theoretic definition of similarity. In:*Proceeding of 15th International Conference of Machine Learning.*.

Lua, K. T. (1993). A study of Chinese word semantics and its prediction. *Computer Processing of Chinese and Oriental Languages*, Vol 7. No 2.

Lua, K.T. (1997). Prediction of meaning of bisyllabic Chinese words using back propagation neural network. In:*Computer Processing of Oriental Languages. 11(2).*

Lua, K. T. (2002). The Semantic Transformation of Chinese Compound Words (漢語造詞過程的語意轉換). *The 3rd workshop on Chinese lexical semantics*, Taipei.

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

Mei et al (1998). 同義詞詞林. Dong-Hua Bookstore: Taipei.