Semantic classification of Chinese unknown words

Huihsin Tseng
Linguistics
University of Colorado
at Boulder
tseng@colorado.edu

Abstract

This paper describes a classifier that assigns semantic thesaurus categories to unknown Chinese words (words not already in the CiLin thesaurus and the Chinese Electronic Dictionary, but in the Sinica Corpus). The focus of the paper differs in two ways from previous research in this particular area.

Prior research in Chinese unknown words mostly focused on proper nouns (Lee 1993, Lee, Lee and Chen 1994, Huang, Hong and Chen 1994, Chen and Chen 2000). This paper does not address proper nouns, focusing rather on common nouns, adjectives, and verbs. My analysis of the Sinica Corpus shows that contrary to expectation, most of unknown words in Chinese are common nouns, adjectives, and verbs rather than proper nouns. Other previous research has focused on features related to unknown word contexts (Caraballo 1999; Roark and Charniak 1998). While context is clearly an important feature, this paper focuses on non-contextual features, which may play a key role for unknown words that occur only once and hence have limited context. The feature I focus on, following Ciaramita (2002), is morphological similarity to words whose semantic category is known. My nearest neighbor approach to lexical acquisition computes the distance between an unknown word and examples from the CiLin thesaurus based upon its morphological structure. The classifier improves on baseline semantic categorization performance for adjectives and verbs, but not for nouns.

1 Introduction

The biggest problem for assigning semantic categories to words lies in the incompleteness of dictionaries. It is impractical to construct a dictionary that will contain all words that may occur in some previously unseen corpora. This issue is particularly problematic for natural language processing applications that work with Chinese texts. Specifically, for the Sinica Corpus[1], Bai, Chen and Chen (1998) found that articles contain on average 3.51% words that were not listed in the Chinese Electronic Dictionary[2] of 80,000 words. Because novel words are created daily, it is impossible to collect them all. Furthermore, across most of the corpora, many of these newly coined words seem to be used only once, and thus they may not even be worth collecting. However, the occurrence of unknown words makes a number of NLP (Natural Language Processing) tasks such as segmentation and word sense disambiguation more difficult. Consequently, it would be valuable to have some means of automatically assigning meaning to unknown words. This paper describes a classifier that assigns semantic thesaurus categories to unknown Chinese words.

The Caraballo (1999)'s system adopted the contextual information to assign nouns to their hyponyms. Roark and Charniak (1998) used the co-occurrence of words as features to classify nouns. While context is clearly an important feature, this paper focuses on non-contextual features, which may play a key role for unknown words that occur only once

[1] The Sinica Corpus is a balanced corpus contained five million part-of-speech words in Mandarin Chinese.

[2] The Chinese Electronic Dictionary is from the Computational Linguistics Society of R.O.C.

and hence have limited context. The feature I focus on, following Ciaramita (2002), is morphological similarity to words whose semantic category is known. Ciaramita (2002) boosted the lexical acquisition system by simple morphological rules and found a significant improvement. Such a finding suggests that a reliable source of semantic information lies in the morphology used to construct the unknown words.

In Chinese morphology, the two ways to generate new words are compounding and affixation. Orthographically, such compounding and affixation is represented by combinations of characters, and as a result, the character combinations and the morpho-syntactic relationship used to link them together can be clues for classification. Furthermore, my analysis of the Sinica Corpus indicates that only 49.68% monosyllabic[3] words have one word class, but 91.67% multisyallabic words have one word class in Table 1. Once characters merge together, only 8.33% words remain ambiguous. It implies that as characters are combined together, the degree of ambiguity tends to decrease.

Word Class[4]	Monosyllabic	Multisyllabic
1	49.68%	91.67%
2	21.94%	7.30%
3	10.94%	0.82%
4	6.55%	0.15%
more than 4	10.89%	0.06%

Table 1 The ambiguity distribution of monosyllabic and multisyllabic words

The remainder of this paper is organized in the following manner: section 2 introduces the CiLin thesaurus, section 3 provides an analysis of unknown words in the Sinica Corpus, and section 4 details the algorithm used for the semantic classification and explains the results.

2 The CiLin thesaurus

The CiLin (Mei et al 1986) is a thesaurus that contains 12 main categories: A-human, B-object, C-time and space, D-abstract, E-attribute, F-action, G-mental action, H-activity, I-state, J-association, K-auxiliary, and L-respect. The majority of words in the A-D categories are nouns, while the majority in the F-J categories are verbs. As shown in Figure 1, the main categories are further subdivided into more specific subcategories in a three-tier hierarchy.

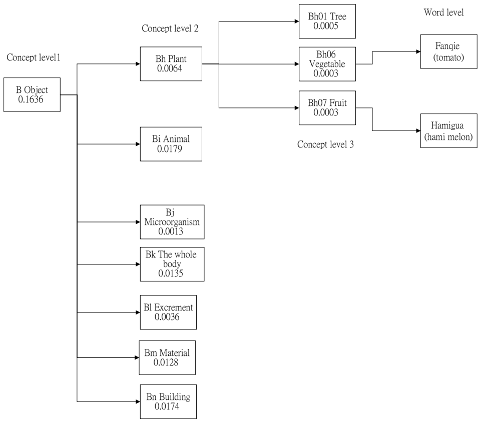

Figure 1 The taxonomy of the CiLin with the probability (partial)

3 Corpus analysis of Chinese unknown words

3.1 Definition of unknown words

Unknown words are the Sinica Corpus lexicons that are not listed in the Chinese Electronic Dictionary of 80,000 lexicons and the CiLin. The 5 million word Sinica Corpus contains 77,866 unknown words consisting of 1.59% adjectives, 33.73% common nouns, 25.18% proper nouns, 12.48% location nouns, 2.98% time nouns, and 24.04% verbs as shown in Table 2.

The focus of most other Chinese unknown word research is on identification of proper nouns such as proper names (Lee 1993), personal names (Lee, Lee and Chen 1994), abbreviation (Huang, Hong and Chen 1994), and organization names (Chen & Chen 2000). Unknown words in categories outside

[3] 'Monosyllabic word' means a word with only a character, and 'multisynllabic word' means a word with more than one character.

[4] 'Word Class' means the number of each word's word class.

the class of proper nouns are seldom mentioned. One of the few examples of multiple class word prediction is Chen, Bai and Chen‘s 1997 work employing statistical methods based on the prefix-category and suffix-category associations to predict the syntactic function of unknown words. Although proper nouns may contain lots of useful and valuable information in a sentence, the majority of unknown words in Chinese are lexical words, and consequently, it is also important to classify lexical words. If not, the remaining 70% of unknown words[5] will be an obstacle to Chinese NLP, where 24.04% of verbs are unknown can be a major problem for parsers.

Class	Unknown words	Corpus lexicons[6]
Adjective	1.59%	1.49%
Common noun	33.73%	37.12%
Proper noun[7]	25.18%	16.53%
Location noun[8]	12.48%	10.38%
Time noun[9]	2.98%	2.36%
Verb	24.04%	32.11%

Table 2 The distribution of unknown words and all lexicons of the Sinica Corpus in 6 classes

3.2 Types of unknown words

In Chinese morphology, the two ways to generate new words are compounding and affixation.

Compounds

A compound is a word made up of other words. In general, Chinese compounds are made up of words that are linked together by morpho-syntactic relations such as modifier-head, verb-object, and so on (Chao 1968, Li and Thompson 1981). For example, 光幻覺/guanghuanjue LIGHT-ILLUSION ‘optical illusion’, consists of 光/guang ‘light’ and 幻覺 /huanjue ‘illusion’, and the relation is modifier-head. 光過敏/ guangguomin LIGHT-ALLERGY ‘photosensitization’ is made up of光/ guang ‘light’ and 過敏/ guomin ‘allergy’, and the relation is modifier-head.

Affixation

A word is formed by affixation when a stem is combined with a prefix or a suffix morpheme. For example English suffixes such as -ian and -ist are used to create words referring to a person with a specialty, such as `musician' and `scientist'. Such suffixes can give very specific evidence for the semantic class of the word. Chinese has suffixes with similar meanings to -ian or -ist, such as the Chinese suffix -jia. But the Chinese affix is a much weaker cue to the semantic category of the word than English -ist or -ian, because it is more ambiguous. The suffix –jia contains three major concepts: 1) expert, such as 科學家/kexuejia SCIENCE-EXPERT ‘scientist’ and 音樂家/ yinyuejia MUSIC-EXPERT ‘musician’, 2) family and home, such as 全家/quanjia WHOLE-FAMILY ‘whole family’ and 富貴家/fuguijia RICH-FAMILY ‘rich family’, 3) house, such as 搬家/banjia MOVE-HOUSE ‘to move house’. In English, the meaning of an unknown word with the suffix –ian or –ist is clear, but in Chinese an unknown word with the suffix –jia could have multiple interpretations. Another example of ambiguous suffix, –xing, has three main concepts: 1) gender, such as 女性/nuxing FEMALE-SEX ‘female’, 2) property, such as 藥性/yaoxing MEDICINE-PROPERTY ‘property of a medicine’, 3) a characteristic, 嗜殺成性/shishachengxing LIKE-KILL-AS-HABIT ‘a characteristic of being bloodthirsty’. Even though Chinese also has morphological suffixes to generate unknown words, they do not determine meaning and syntactic category as clearly as they do in English.

[5] Part of location noun still contains some proper nouns like country names.

[6] It contains both known and unknown words.

[7] Proper noun contains two classes: 1) formal name, such as personal names, races, titles of magazines and so on. 2) Family name, such as Chen and Lee.

[8] Location noun contains 4 subclasses: 1) country names, such as China. 2) common location noun, such as 郵局 /youju ‘post office’ and 學校/xuexiao ‘school’. 3) noun + position, such as 海外/haiwei ‘oversea’. 4) direction noun, such as 上/shang ‘up’ and 下/xia ‘down’.

[9] Time noun contains 3 classes: 1) historical event and recursive time noun, such as 清/Qing dynasty and 一月 /yiyue ‘January’. 2) noun + position, such as 晚間 /wanjian ‘in the evening’, 3) adverbial time noun, such as 將來/jianglai ‘in the future’.

4 Semantic classification

For the task of classifying unknown words, two algorithms are evaluated. The first algorithm uses a simple heuristic where the semantic category of an unknown word is determined by the head of the unknown word. The second algorithm adopts a more sophisticated nearest neighbor approach such that the distance between an unknown word and examples from the CiLin thesaurus computed based upon its morphological structure. The first algorithm serves to provide a baseline against which the performance of the second can be evaluated.

4.1 Baseline

The baseline method is to assign the semantic category of the morphological head to each word.

4.2 An example-base semantic classification

The algorithm for the nearest neighbor classifier is as follows:

1) An unknown word is parsed by a morphological analyzer (Tseng and Chen 2002). The analyzer a) segments a word into a sequence of morphemes, b) tags the syntactic categories of morphemes, and c) predicts morpho-syntactic relationships between morphemes, such as modifier-head, verb-object and resultative verbs as shown as in Table 3. For example, if 舞蹈家/wudaojia DANCE-EXPERT ‘dancer’ is an unknown word, the morphological segmentation is 舞蹈/wudao DANCE ‘dance’ and 家/jia EXPERT ‘expert’, and the relation is modifier-head.

2) The CiLin thesaurus is then searched for entries (examples) that are similar to the unknown word. A list of words sharing at least one morpheme with the unknown word, in the same position, is constructed. In the case of 舞蹈家/wudaojia, such a list would include 歌唱家/gechangjia SING-EXPERT ‘singer’, 回家/huijia GO-HOME ‘go home’, 富貴家/fuguijia RICH-FAMILY ‘rich family’ and so on.

Word Class	The morpho-syntactic relations
Noun	Modifier-head[10] 籃球/lanqie BASKET-BALL `baseketball’
Verb	1) Verb-object : 吃飯/chifan EAT-RICE ‘to eat` 2) Modifier-head: 清列/qinglie CLEAR-LIST ‘clearly list’ 3) Resultative Verb 吃飽/chibao EAT-FULL ‘to have eaten’ 4) Head-suffix: 變成/biancheng CHANG-TO ‘become’ 5) Modifier-head (suffix): 自動化/zidonghua AUTOMATIC-BECOME ‘automatize’ 6) Directional resultative compound and reduplication 跑上來/paoshanglai RUN-UP-TO ‘run up to’
Adjective	An: modifier-head 中國式/zhongguoshi CHINESE-STYLE ‘Chinese stylish’ Av: verb-object and modifier-head 愚民/yumin FOOL-PEOPLE ‘keeping the people uninformed’

Table 3 The morpho-syntactic relations

3) The examples that do not have the same morpho-syntactic relationships but shared morpheme belongs to the unknown word’s modifier are pruned away. If no examples are found, the system falls back to the baseline classification method.

4) The semantic similarity metric used to compute the distance between the unknown word and the selected examples from the CiLin thesaurus is based upon a method first proposed by Chen and Chen (1997).

They assume that similarity of two semantic categories is the information content of their parent’s

[10] There are still a very small number of coordinate relation compounds that is both of the morphemes in a compound are heads. Since either one of the morphemes can be the meaning of the whole compound, in order to simplify the system, words that have coordinate relations are categorized as modifier head relation.

node. For instance, the similarity of 哈密瓜/hamigua ‘hami melon’ (Bh07) and 番茄/fanqie ‘tomato’ (Bh06) is based on the information content of the node of their least common ancestor Bh.

The CiLin thesaurus can be used as an information system, and the information content of each semantic category is defined as

$$\text{Entropy(System)} - \text{Entropy(Semantic category)}$$

The similarity of two words is the least common ancestor information content(IC), and hence, the higher the information content is, the more similar two the words are. The information content is normalized by Entropy(system) in order to keep the similarity between 0 and 1. To simplify the computation, the probabilities of all leaf nodes are assumed equal. For example, the probability of Bh is .0064 and the information content of Bh is –log(.0064). Hence, the similarity between 哈密瓜/ hamigua and 番茄/ fanqie is .61.

$$\mathrm{Sim}\left(W_1 \cap W_2\right) = \frac{\mathrm{IC}\left(W_1 \cap W_2\right)}{\mathrm{Entropy}\left(\mathrm{System}\right)} = \frac{-\log_2\left(\mathrm{P}\left(W_1 \cap W_2\right)\right)}{\mathrm{Entropy}\left(\mathrm{System}\right)} \qquad (1)$$

Let System = CiLin,
W_1 = Bh07 (the category of hamihua),
W_2 = Bh06 (the category of fanqie)

$$\mathrm{Sim}\left(\mathrm{Bh07} \cap \mathrm{Bh06}\right) = \frac{\mathrm{IC}\left(\mathrm{Bh07} \cap \mathrm{Bh06}\right)}{\mathrm{Entropy}\left(\mathrm{CiLin}\right)} = \frac{-\log_2\left(\mathrm{P}\left(\mathrm{Bh}\right)\right)}{\mathrm{Entropy}\left(\mathrm{CiLin}\right)}$$
$$= \frac{-\log_2 0.0064}{-\log_2 0.0026} = \frac{7.29}{11.94} = 0.61$$

Resnik (1995, 1998 and 2000) and Lin (1998) also proposed information content algorithms for similarity measurement. The Chen and Chen (1997) algorithm is a simplification of the Resnik algorithm, which makes the simplifying assumption that the occurrence probability of each leaf node is equal.

One problem for this algorithm is the insufficient coverage of the CiLin (CiLin may not cover all morphemes). The backup method is to run the classifier recursively to predict the possible categories of the unlisted morphemes. If a morpheme of an unknown word or of an unknown word’s example is not listed in the CiLin, the similarity measurement will suspend measuring the similarity between the unknown word and the examples and run the classifier to predict he semantic category of the morpheme first. After the category of the morpheme is known, the classifier will continue to measure the similarity between the unknown word and its examples. The probability of adopting this backup method in my experiment is on the average of 3%.

Here is an example of the recursive semantic measurement. 跑碼頭/paomatou RUN-WHARF ‘wharf-worker’ is an example of an unknown word 跑旱船/paohanchuan RUN-DRY BOAT ‘folk activities’. The morphological analyzer breaks the two words into 跑 碼頭/pao matou and 跑 旱船/pao hanchuan. The measurement function will compute the similarity between 碼頭/matou and 旱船/hanchuan, but in this case, 旱船/hanchuan is not listed in the CiLin. The next approach is then to run the semantic classifier to guess the possible category of 旱船/hanchuan. Based on the predicted category, it then goes on to compute the similarity for 碼頭/matuo and 旱船/hanchuan. By applying this method, there will not be any words without a similarity measurement.

5) After the distances from the unknown word to each of the selected examples from the CiLin thesaurus are determined, the average distance to the *K* nearest neighbors from each semantic category is computed. The category with the lowest distance is assigned to the unknown word.

The similarity of 舞蹈/wudao and 歌唱/gechang is .87, of 舞蹈/wudao and 回/hui is .26, and of 舞蹈/wudao and 富貴/fugui is 0. Thus, 舞蹈家/wudaojia is more similar to 歌唱家/gechangjia than回家/huijia or富貴家/fuguijia. The category of 舞蹈家/wudaojia is thus most likely to be 歌唱家/gechangjia.

The semantic category is predicted as the category that gets the highest score in formula (2). The lexical similarity and frequency of examples of each category are considered as the most important features to decide a category.

In formula (2), RankScore(C_i) includes SS(C_i) and FS(C_i). The score of SS(C_i) is a lexical similarity score, which is from the maximum score of Similarity (W_1,W_2) in the category of W_2. FS(C_i) is a frequency score to show how many examples there are in a category. α and (1-α) are respectively weights for the lexical similarity score and the frequency score.

$$\begin{aligned}\text{Let } W_1 &= \text{unknownword}\\ W_i &= \text{word whose semantic category defined in the CiLin}\\ i &= \text{A...L(CiLin Taxonomy)}\end{aligned}$$

$$\text{Rankscore}(C_i) = \alpha * \text{SS}(C_i) + (1-\alpha) * \text{FS}(C_i) \quad (2)$$

$$\text{SS}(C_i) = \underset{C(W_i)\in C_i}{\overset{i=\text{A...L}}{\arg\max}}\, \text{Sim}(W_1, W_i) \quad (3)$$

$$\text{FS}(C_i) = \frac{\text{Freq}(C_i)}{\sum_{i=\text{A}}^{\text{L}} \text{Freq}(C_i)} \quad (4)$$

5 Experiment

5.1 Data

There are 56,830 words in the CiLin. For experiments, CiLin lexicons are divided into 2 sets: a training set of 80% CiLin words, a development set of 10% of CiLin words, and a test set of 10% CiLin words. All words in the test set are assumed to be unknown, which means the semantic categories in both sets are unknown. Nevertheless, the morphological structures of proper nouns are different from lexical words. Their identification methods are also different and will be out of the scope of this paper. The correct category of the unknown word is the semantic category in the CiLin, and if an unknown word is ambiguous, which means it contains more than one category, the system then chooses only one possible category. In evaluation, any one of the categories of an ambiguous word is considered correct.

5.2 Result

On the test set, the baseline predicts 53.50% of adjectives, 70.84% of nouns and 47.19% of verbs correctly. The classifier reaches 64.20% in adjectives, 71.77% in nouns and 53.47% in verbs, when α is 0.5 and K is five.

Word class	Baseline accuracy	Semantic classification accuracy
Adjective	53.50%	64.20%
Noun	70.84%	71.77%
Verb	47.19%	53.47%

Table 4 The accuracy of the baseline and semantic classification in the development set

Word class	Baseline accuracy	Semantic classification accuracy
Adjective	52.92%	65.76%
Noun	70.89%	71.39%
Verb	44.10%	52.84%

Table 5 The accuracy of the baseline and semantic classification in the test set

Table 4 and table 5 show a comparison of the baseline and the classifier. Generally, nouns are easier to predict than the other categories, because their morpho-syntactic relation is not as complex as verbs and adjectives. The classifier improves on baseline semantic categorization performance for adjectives and verbs, but not for nouns. The lack of a performance increase for nouns is most likely because nouns only have one kind of morpho-syntactic relation. The advantage of the classifier is to filter out examples in different relations and to find out the most similar example in morphemes and morpho-syntactic relation. The classifier predicts better than the baseline in word classes with multiple relations, such as adjectives and verbs. For example, 開快車/kaikuaiche OPEN-FAST CAR 'drive fast' is a verb-object verb. The baseline wrongly predicted it due to the verb, 開/kai OPEN 'open'. However, the semantic classifier grouped it to the category of its similar example, 開夜車/kaiyeche OPEN-NIGHT CAR 'drive during the night'.

5.3 Error analysis

Error sources can be grouped into two types: data errors and the classifier errors. The testing data is from the CiLin. Some of testing data are not semantically transparent such as idioms, metaphors, and slang. The meaning of such words is different from the literal meaning. For instance, the literal meaning of 看門狗/kanmengou WATCH-DOOR-

DOG is a door-watching dog, and in fact it refers to a person with the belittling meaning. 母老虎/mulaohu FEMALE-TIGER is a female tiger literally, and it refers to a mean woman. These words do not carry the meaning of their head anymore. An unknown word will be created such as 看門貓/kanmenmao WATCH-DOOR-CAT ‘a door-watching cat’, but it is impossible for unknown words to carry similar meaning of words as 看門狗/kanmengou.

The classifier errors are due primarily to three factors: a lack of examples, the preciseness of the similarity measurement, and the taxonomy of the CiLin.

First, some errors occur when there are not enough examples in training data. For example, 鐵欄杆/tielangan IRON-POLE ‘iron pole` does not have any similar examples after the classifier filters out examples whose relations are different and whose shared morphemes are not head. 鐵欄杆/tielangan is segmented as 鐵/tie IRON ‘iron’ and 欄杆/langan POLE ‘pole’. There are examples of the first morpheme, 鐵/tie, but no similar examples of the second,欄杆/langan. Since 鐵欄杆/tielangan has modifier-head relation and 欄杆/langan is the head of the compound, then the classifier filters out the examples of鐵/tie. There are hence not enough examples. Filtering examples in different structures is performed to make the remaining examples more similar since the similarity measurement may not be able to distinguish slight differences. However, the cost of this filtering of different structure examples is that sometimes this leaves no examples.

Second, the similarity measurement is sometimes not powerful enough. 運動場/yundongchang SPORT-SPACE ‘a sports ground` has a sufficient number of examples, but has problems with the similarity measurement. The head 場/chang is ambiguous. 場/chang has two senses and both mean space. One of them means abstract space and the other means physical space. Hence, in the CiLin thesaurus 場/chang can be found in C (time and space) and D (abstract). Words in C such as 商場/shangchang BUSINESS-SPACE ‘a market’, 屠宰場/tuzaichang BUTCHER-SPACE ‘a slaughter house’ , 會場/huichang MEETING-SPACE ‘the place of a meeting’, and in D are 球場/ qiuchang BALL-SPACE ‘a court’, 體育場/tiyuchang PHYSICAL TRAINING-SPACE ‘a stadium’. 運動場/yundongchang should be more similar to 體育場/tiyuchang than other space nouns, but the similarity score does not show that they are related and C group has more examples. Thus, the system chooses C incorrectly.

Third, the taxonomy of the thesaurus is ambiguous. For instance, 體操房/tichaofang GYMNASTICS–ROOM ‘gymnastics room’ has similar examples in both B (object) and D (abstract). These two groups are very similar. Words in B group include 刑房/xingfan PUNISHMENT-ROOM ‘punishment room’, 書房/shufan BOOK-ROOM ‘study room’, 暗房/anfan DARK-ROOM ‘dark room’, and 廚房/chufan KITCHEN-ROOM ‘kitchen’. Words in D are such as 牢房/laofan PRISON-ROOM ‘a jail’ and 彈子房/danzifan BILLIARD-ROOM ‘a billiard room’. There are no obvious features to distinguish between these examples. According to the CiLin, 體操房/tichaofang belongs to D, but the classifier predicts it as B class which does not actually differ much with D. Such problems may occur with any semantic taxonomy.

6 Conclusion

The paper presents an algorithm for classifying the unknown words semantically. The classifier adopts a nearest neighbor approach such that the distance between an unknown word and examples from the CiLin thesaurus is computed based upon its morphological structure. The main contributions of the system are: first, it is the first attempt in adding semantic knowledge to Chinese unknown words. Since over 70% of unknown words are lexical words, the inability to resolve their meaning is a major obstacle to Chinese NLP such as semantic parsers. Second, without contextual information, the system can still successfully classify 65.76% of adjectives, 71.39% of nouns and 52.84% of verbs. Future work will explore the use of the contextual information of the unknown words and the contextual information of the lexicons in the predicted category of the unknown words to boost predictive power.

Acknowledgment

Thanks to S. Bethard, D. Cer, K. J. Chen, D. Jurafsky and to the anonymous reviewers for many helpful suggestions. This research was partially supported by the NSF via a KDD extension to NSF IIS-9978025 (Dan Jurafsky, PI) and by the CKIP group, Institute of Information Science, Academia Sinica.

References

Bai, M. H., C.J. Chen, and K. J. Chen. 1998. “白明弘、陳超然、陳克健”。 <以語境判定中文未知詞詞類的方法>，«第十一屆計算機語言學會論文集»。頁 47-60。

Caraballo, S. 1999. Automatic acquisition of a hypemymlabeled noun hierarchy from text, in Proceedings of the 37th ACL.

Ciaramita. M. 2002. Boosting automatic lexical acquisition with morphological information", in Proceedings of the Workshop on Unsupervised Lexical Acquisition, ACL-02.

Chao, Y. R. 1968. A grammar of spoken Chinese. Berkeley:University of California Press.

Chen, C. J., M. H. Bai and K. J. Chen. 1997. Category Guessing for Chinese Unknown Words, in Proceedings of the Natural Language Processing Pacific Rim Symposium, 35-40.

Lee, J. C. 1993. “李振昌” 。«中文文本專有名詞辨識問題之研究»。臺灣大學資訊工程研究所碩士論文。

Lee, J. C., Y. H. Lee and H. H. Chen. 1994. “李振昌、李御璽、陳信希”。«中文文本人名辨識問題之研究»。<第七屆計算器語言會會議論文集>，頁 203-222。

Chen. K. J. and C. J Chen. 1997. “陳克健、陳超然”。<語料庫為本的中文複合詞構詞律模型研究>，«漢語計量與計算研究»，編輯：鄒嘉彥、黎邦洋、陳偉光、王士元。頁 283-305。香港：城市大學。

Chen, K. J. and M. H. Bai. 1998. Unknown Word Detection for Chinese by a Corpus-based Learning Method, in Computational Linguistics and Chinese Language Processing vol3 no. 1, 27-44.

Chen, C. J. and K. J. Chen. 2002. Knowledge Extraction for Identification of Chinese Organization Names, in Proceedings of the second Chinese Language Processing Workshop, 15-21.

Huang, C. R., W. M. Hong and K. J. Chen. 1994. An Introduction Based Lexical of Abbreviation, in Proceedings of the 2th Pacific Asia Conference on Formal and Computational Linguistics, 49-52.

Huang, C. R. and K. J. Chen. 1995. “黃居仁 陳克健” 。«中央研究院平衡語料庫»。中研院詞庫小組。

Li, C. and S. A. Thompson. 1981. Mandarin Chinese. Berkeley: University of California Press.

Lin, D.. 1998. An information-theoretic definition of similarity, in Proceedings 15th International Conf. on Machine Learning, p 296—304.

Lin, D. and P. Pantel.. 2001. Induction of Semantic Classes from Natural Language Text, In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001, 317-322.

Mei, J., Y. Zhu., Y. Gao, and H. Ying. 1986. “梅家駒、竺一鳴、高蘊琦、殷鴻翔”。1986。«同義詞詞林»。香港：商務印書館。”

Resnik, P.. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448-453.

---. 1998. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, in Journal of Artificial Intelligence Research (11), 95-130.

Resnik, P. and M. Diab. 2000. Measuring Verbal Similarity. Technical Report: LAMP-TR-047//UMIACS-TR-2000-40/CS-TR-4149/MDA-9049-6C-1250. University of Maryland, College Park.

Roark, B. and E. Charniak. 1998. Noun-phrase co-occurrence statistics from semi-automatic semantic lexicon construction, in Proceedins of the 36th ACL.

Tseng, H and K. J. Chen. 2002. Design of Chinese Morphological Analyzer. SigHan Workshop on Chinese Language Processing, Taipei.