Extending a Thesaurus with Words from Pan-Chinese Sources

Oi Yee Kwong[†‡] **and Benjamin K. Tsou**[‡]
[†]Department of Chinese, Translation and Linguistics
[‡]Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
{rlolivia, rlbtsou}@cityu.edu.hk

Abstract

In this paper, we work on extending a Chinese thesaurus with words distinctly used in various Chinese communities. The acquisition and classification of such region-specific lexical items is an important step toward the larger goal of constructing a Pan-Chinese lexical resource. In particular, we extend a previous study in three respects: (1) to improve automatic classification by removing duplicated words from the thesaurus, (2) to experiment with classifying words at the subclass level and semantic head level, and (3) to further investigate the possible effects of data heterogeneity between the region-specific words and words in the thesaurus on classification performance. Automatic classification was based on the similarity between a target word and individual categories of words in the thesaurus, measured by the cosine function. Experiments were done on 120 target words from four regions. The automatic classification results were evaluated against a gold standard obtained from human judgements. In general accuracy reached 80% or more with the top 10 (out of 80+) and top 100 (out of 1,300+) candidates considered at the subclass level and semantic head level respectively, provided that the appropriate data sources were used.

1 Introduction

A unique problem in Chinese language processing arises from the extensive lexical variations among major Chinese speech communities. Although different communities (e.g. Beijing, Hong Kong, Taipei and Singapore) often share a large core lexicon, lexical variations could occur in at least two ways. On the one hand, even the same word forms shared by various communities could be used with different meanings. For instance, the word 居屋 (*ju1wu1*)[1] refers to general housing in Mainland China but specifically to housing under the Home Ownership Scheme in Hong Kong. On the other hand, there are substantially different lexical items used for lexicalizing common or region-specific concepts. For example, while the word 住房 (*zhu4fang2*) is similarly used as 居屋 to mean general housing in Mainland China, it is rarely seen in the Hong Kong context; and 下崗 (*xia4gang3*) is specific, if not exclusive, to Mainland China for referring to a special concept of unemployment.

Existing Chinese lexical resources are often based on language use in one particular region and are therefore not comprehensive enough to capture the substantial *regional variation* as an important part of the lexical knowledge, which will be useful and critical for many NLP applications, including natural language understanding, information retrieval, and machine translation.

Tsou and Kwong (2006) proposed a comprehensive Pan-Chinese lexical resource, using a large and unique synchronous Chinese corpus as an authentic source of lexical variation among various Chinese speech communities. They also studied the feasibility of taking an existing Chinese thesaurus as leverage and classifying new words from various Chinese communities with respect to the classificatory structure therein (Kwong and Tsou, 2007). They used the catego-

[1] The transcriptions in brackets are based on Hanyu Pinyin.

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 457–464
Manchester, August 2008

ries at the subclass level of the *Tongyici Cilin* (同義詞詞林, abbreviated as Cilin hereafter) for the task. The classification was done by comparing the similarity of a target word (i.e. the word to be classified) and individual categories of words in the thesaurus based on a feature vector of co-occurring words in a corpus. Since words in the thesaurus are mostly based on lexical items used in Mainland China, and the target words come from various Chinese communities, a major issue in the classification task is thus the heterogeneity of the data sources. It was hypothesized that the datasets from which the features were extracted (for the target words and words in the thesaurus respectively) may affect the performance of automatic classification. The experimental results supported the hypothesis in part, and the actual effect varied with datasets from individual regions. Moreover, there is room to improve the overall accuracy for the method to be useful in practice, and it appears that the duplicated words in the thesaurus might have skewed the similarity measurement to a certain extent.

The current study thus attempts to extend this previous study in three respects: (1) to improve automatic classification by removing duplicated words from the thesaurus, (2) to experiment with classifying words at the subclass level and semantic head level (a finer level), and (3) to further investigate the possible effects of data heterogeneity between the region-specific words and words in the thesaurus on classification performance.

In Section 2, we will briefly review related work and the background of the current study. In Sections 3 and 4, we will describe the materials used and the experimental setup respectively. Results will be presented in Section 5 and discussed in Section 6, followed by a conclusion in Section 7.

2 Related Work

To build a semantic lexicon, one has to identify the relation between words within a semantic hierarchy, and to group similar words together into a class. Previous work on automatic methods for building semantic lexicons could be divided into two main groups. One is automatic thesaurus acquisition, that is, to identify synonyms or topically related words from corpora based on various measures of similarity (e.g. Riloff and Shepherd, 1997; Lin, 1998; Caraballo, 1999; Thelen and Riloff, 2002; You and Chen, 2006).

Another line of research, which is more closely related to the current study, is to extend existing thesauri by classifying new words with respect to their given structures (e.g. Tokunaga *et al.*, 1997; Pekar, 2004). An early effort along this line is Hearst (1992), who attempted to identify hyponyms from large text corpora, based on a set of lexico-syntactic patterns, to augment and critique the content of WordNet. Ciaramita (2002) compared several models in classifying nouns with respect to a simplified version of WordNet and signified the gain in performance with morphological features. For Chinese, Tseng (2003) proposed a method based on morphological similarity to assign a Cilin category to unknown words from the Sinica corpus which were not in the Chinese Electronic Dictionary and Cilin; but somehow the test data were taken from Cilin, and therefore could not really demonstrate the effectiveness with unknown words found in the Sinica corpus.

Kwong and Tsou (2007) attempted to classify words distinctly used in Beijing, Hong Kong, Singapore, and Taiwan, with respect to the Cilin classificatory structure. They brought up the issue of data heterogeneity in the task. In general, automatic classification of words via similarity measurement between two words, or between a word and a class of words, was often done on words from a similar data source, with the assumption that the feature vectors under comparison are directly comparable. In the Pan-Chinese context, however, the words to be classified come from corpora collected from various Chinese speech communities, but the words in the thesaurus are often based on usages found in a particular community, such as Mainland China in the case of Cilin. It is thus questionable whether the words in Cilin would appear in comparable contexts in texts from other places, thus affecting the similarity measurement. In view of this heterogeneous nature of the data, they experimented with extracting feature vectors for the Cilin words from different datasets and found that the classification of words from Taipei was most affected in this regard.

In general, up to 85% accuracy was reached with the top 15 candidates for classification at the Cilin subclass level. This performance, however, should be improved for the method to be useful in practice. It is observed that Cilin, as most other thesauri, does not have a mutually exclusive classification. Many words appear in more than one category (at various levels). Such duplication may affect the similarity comparison

between a target word and words in a category. The current study thus attempts to avoid this confounding factor by removing duplicated words from Cilin for the comparison of similarity, and to extend the classification to a finer level.

3 Materials

3.1 The Tongyici Cilin

The *Tongyici Cilin* (同義詞詞林) (Mei *et al.*, 1984) is a Chinese synonym dictionary, or more often known as a Chinese thesaurus in the tradition of the Roget's Thesaurus for English. The Roget's Thesaurus has about 1,000 numbered semantic heads, more generally grouped under higher level semantic classes and subclasses, and more specifically differentiated into paragraphs and semicolon-separated word groups. Similarly, some 70,000 Chinese lexical items are organized into a hierarchy of broad conceptual categories in Cilin. Its classification consists of 12 top-level semantic classes, 94 subclasses, 1,428 semantic heads and 3,925 paragraphs. It was first published in the 1980s and was based on lexical usages mostly of post-1949 Mainland China. In the current study, we will focus on the subclass level and semantic head level. Some example subclasses and semantic heads are shown in Table 1.

We classify words with respect to the subclass level and semantic head level (that is, second and third levels in the Cilin organisation). Moreover, we skip class K and class L as the former contains mostly function words and the latter longer expressions. We are thus considering 88 subclasses and 1,356 semantic heads in this study.

Within classes A to J, there are 7,517 words which were found to appear in more than one category. Upon removing these entries, 44,588 words were used in the similarity comparison for the current study.

Class	Subclasses	Semantic Heads
A 人 (Human)	Aa … Ae 職業 (Occupation) Af 身份 (Identity) … An	Aa01 … Ae10 軍官 將士 軍人 士兵 (commander, soldier) … An07
B 物 (Things)	Ba … Bb 擬狀物 (Shape) … Bi 動物 (Animal)… Bm 材料 (Material)…Bq 衣物 (Clothing) … Br	Ba01 … Bm08 煤 炭 (coal, carbon) … Bn03 房間 (room) … Br14
C 時間與空間 (Time and Space)	Ca 時間 (Time) Cb 空間 (Space)	Ca01 … Ca18 年 (year) … Cb28 場所 (location) … Cb30
D 抽象事物 (Abstract entities)	Da 事情 情況 (Condition) … Df 意識 (Ideology) … Di 社會 政法 (Society) Dj 經濟 (Economics) … Dm 機構 (Organization) Dn 數量 單位 (Quantity)	Da01 … Di10 團體 派別 (group, party) … Dj04 資本 利潤 利息 債務 (capital, interest) Dj05 貨幣 票據 (currency, invoice) … Dm01 政府 (government) … Dn10
E 特徵 (Characteristics)	Ea … Ed 性質 (Property)… Ef	Ea01 … Ed03 好 壞 (goodness, badness) … Ef14
F 動作 (Action)	Fa … Fd 全身動作 (Body action)	Fa01 … Fb01 走 跑 (run) … Fd09
G 心理活動 (Psychological activities)	Ga … Gb 心理活動 (Psychological activities)… Gc	Ga01 … Gb01 想像 思考 斟酌 著想 (imagine, think) … Gc04
H 活動 (Activities)	Ha … He 經濟活動 (Economic activities) … Hd 生產 (Production) … Hf 交通 運輸 (Transportation) Hg 教衛科研 (Scientific research)… Hi 社交 (Social contact) Hj 生活 (Livelihood)	Ha01 … Hc09 主持 指揮 統率 執掌 管轄 (in charge, administer, lead) … He03 買 賣 (buy, sell) … Hg01 教育 傳授 示範 (teach, demo) … Hj12 做 施展 合作 嘗試 (do, cooperate, try) … Hn13
I 現象與狀態 (Phenomenon and state)	Ia … If 境遇 (Circumstance) Ig 始末 (Process)… Ih	Ia01 … Ig01 開始 結束 (begin, end) … Ih05 增多 添補 減少 (increase, decrease) … Ih13
J 關聯 (Association)	Ja 聯繫 (Liaison) Jb 異同 (Similarity and Difference) Jc 配合 (Matching) … Je	Ja01 … Jc01 適應 相配 符合 (adapt, match) … Je14

Table 1 Some Examples of Cilin Subclasses and Semantic Heads

3.2 The LIVAC Synchronous Corpus

LIVAC (http://www.livac.org) stands for Linguistic Variation in Chinese Speech Communities. It is a synchronous corpus developed and dynamically maintained by the Language Information Sciences Research Centre of the City University of Hong Kong since 1995 (Tsou and Lai, 2003). The corpus consists of newspaper articles collected regularly and synchronously from six Chinese speech communities, namely Hong Kong, Beijing, Taipei, Singapore, Shang-

hai, and Macau. Texts collected cover a variety of domains, including front page news stories, local news, international news, editorials, sports news, entertainment news, and financial news. Up to December 2007, the corpus has already accumulated over 250 million character tokens which, upon automatic word segmentation and manual verification, yielded about 1.2 million word types.

For the present study, we made use of subcorpora consisting of the *financial news* sections collected over the 9-year period 1995-2004 from Beijing (BJ), Hong Kong (HK), Singapore (SG), and Taipei (TW). Table 2 shows the sizes of the subcorpora.

Region	**Size of Financial Subcorpus** (rounded to nearest 1K)	
	Word Token	**Word Type**
BJ	232K	20K
HK	970K	38K
SG	621K	28K
TW	254K	22K

Table 2 Sizes of Individual Subcorpora

3.3 Test Data

Kwong and Tsou (2006) observed that among the unique lexical items found from the individual subcorpora, only about 30-40% are covered by Cilin, but not necessarily in the expected senses. In other words, Cilin could in fact be enriched with over 60% of the unique items from various regions.

In the current study, we sampled the most frequent 30 words distinctly and predominantly used in each of the BJ, HK, SG, and TW subcorpus. Classification was based on their similarity with each of the Cilin subclasses and semantic heads, compared by the cosine measure, as discussed in Section 4.2.

4 Experiments

4.1 Setting the Gold Standard

Three linguistics undergraduate students and one research student on computational linguistics from the City University of Hong Kong were asked to assign what they would consider to be the most appropriate Cilin category (at the subclass and semantic head level) to each of the 120 target words.

All human judges reported difficulties in various degrees in assigning Cilin categories to the target words. The major problem came from the regional specificity and thus the unfamiliarity of the judges with the respective lexical items and contexts. For example, all judges reported problem with the term 自撮 (*zi4cuo1*), one of the target words from Singapore referring to 自撮股市 (*zi4cuo1gu3shi4*, CLOB in the Singaporean stock market), which is specific to Singapore.

Notwithstanding the difficulty, the inter-annotator agreement, as measured by *Kappa*, was found to be 0.6870 at the subclass level and 0.5971 at the semantic head level.

We took a "loose" approach to form the gold standard, which includes all categories (at the subclass level and semantic head level respectively) assigned by one or more judges. Automatic classification will be considered "correct" if any of these categories is matched.

4.2 Automatic Classification

Each target word was compared to all Cilin categories and automatically classified to the category which is most similar to it. The Cilin data was first pre-processed to remove duplicated words.

We compute the similarity by the cosine between the two corresponding feature vectors containing *all co-occurring content words* in a corpus within a window of ±5 words (excluding many general adjectives and adverbs, and numbers and proper names were all ignored). The feature vector of a Cilin category is based on the union of the features from all individual members in the category.

The cosine of two feature vectors $\vec{v}$ and $\vec{w}$ is computed as

$$\cos(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|}$$

The feature vector of a given target word is extracted from the respective subcorpus from which the target word was found (called the target subcorpus hereafter). To study the data heterogeneity effect, we experimented with two conditions for the extraction of feature vectors for Cilin words: from the target subcorpus or from the BJ subcorpus which is assumed to be representative of usages in Mainland China.

All automatic classification results were evaluated against the gold standard based on hu-

man judgements as discussed in Section 4.1. Classification performance is measured based on the correctness of the top *N* candidates.

4.3 Baseline

A simple baseline measure was obtained by ranking the subclasses in descending order of the number of words they cover. It was assumed that the bigger the subclass size, the more likely it covers a new term. The top *N* candidates in this ranking were checked against the gold standard as above.

5 Results

In the following discussion, we will use labels in the form of <*Cat*>-<*Target*>-<*CilinFeatSource*> to refer to the various testing conditions, where *Cat* refers to the category type, *Target* to the originating source of the target words, and *CilinFeatSource* to the source from which the feature vectors for the Cilin words were extracted. Thus the label Sub-hk-hk means classification of HK target words at the Cilin subclass level, with feature vectors for target words and Cilin words extracted from the HK subcorpus; and the label Head-tw-bj means classification of TW target words at the Cilin semantic head level, with feature vectors for the target words extracted from the TW subcorpus and those for the Cilin words extracted from the BJ subcorpus.

5.1 Pre-processing of Cilin

Figure 1 shows the comparison of classification accuracy for words from the four regions at the subclass level before and after duplicates in Cilin were removed. All feature vectors were extracted from the respective target corpora.

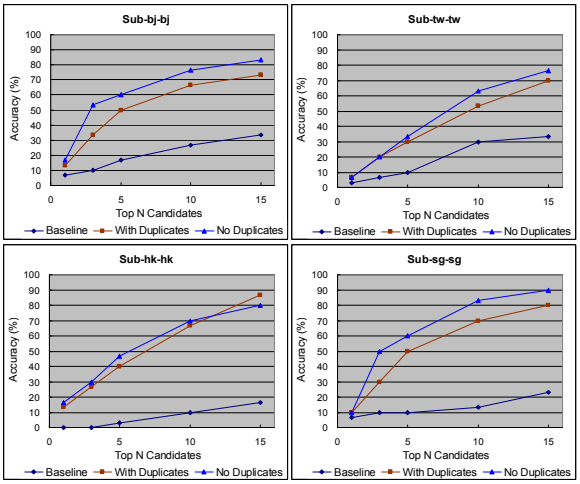

Figure 1 Effect of Pre-processing Cilin

It can be seen from Figure 1 that removing duplicated words in Cilin could improve the classification of words from all regions at the subclass level.

5.2 Data Heterogeneity Effect

As explained earlier, since the words to be classified come from various Chinese speech communities, but the words in Cilin are mostly based on usages found in Mainland China, it is uncertain whether the words in Cilin would appear in comparable contexts in texts from other places, for the similarity measurement to be effective. Hence, we experimented with two conditions for extracting feature vectors for the Cilin words. While the features for a target word to be classified are extracted from the respective target subcorpus, the features for the Cilin words are extracted either from the target subcorpus or from the BJ subcorpus. Figure 2 shows the data heterogeneity effect on the classification of target words from various regions at the subclass level.

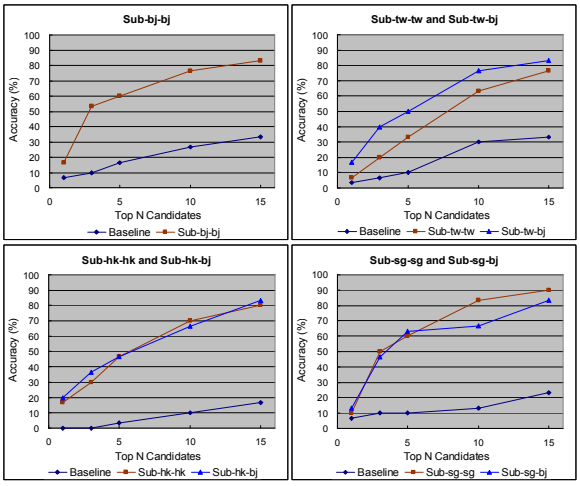

Figure 2 Data Heterogeneity Effect

The data heterogeneity effect is most noticeable for the TW words. Extracting features for the Cilin words from the BJ subcorpus always gives better classification results for the TW words, than if the features were extracted from the TW subcorpus. The difference between Sub-hk-hk and Sub-hk-bj, and that between Sub-sg-sg and Sub-sg-bj, however, is not as great. This suggests that the lexical difference is particularly significant between BJ and TW.

5.3 Fine-grainedness of Classification

The semantic head level is more fine-grained than the subclass level, and is expected to be

more difficult for classification. Figure 3 shows the results of classification at the semantic head level, with the effect of data heterogeneity.

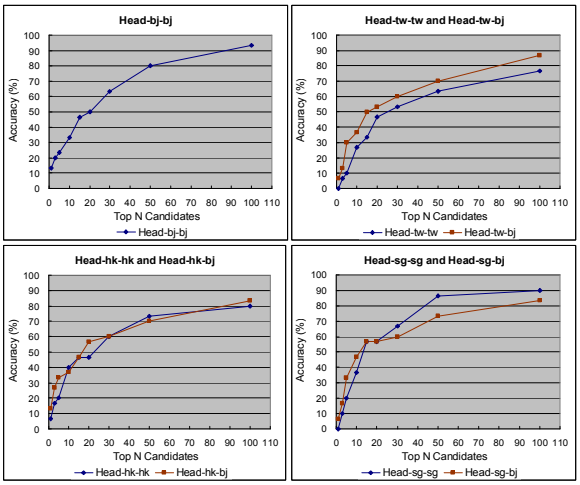

Figure 3 Semantic Head Level Classification

It is observed from Figures 2 and 3 that data heterogeneity affects the classification of TW words at both the subclass and semantic head level. In both cases, features for Cilin words extracted from BJ subcorpus work better than those from TW subcorpus. A somewhat opposite effect was observed for SG target words, especially beyond the top 5 to 10 candidates. There is not much difference for the HK target words.

The classification at the semantic head level is expectedly less precise than that at the subclass level. At the subclass level, 80% or more accuracy could be reached with the top 10 candidates considered, whereas the top 50 candidates or more would be needed to reach a similar level of accuracy at the semantic head level. This is nevertheless encouraging in view of the total number of categories at the semantic head level.

6 Discussions

6.1 Overall Classification Accuracy

From the results reported in the last section, it can be seen that removing the duplicated words in Cilin could help improve the classification accuracy at all conditions. This is because some words, which appear in more than one category at the subclass or semantic head level, might skew the similarity measured between a target word and a given category. An example will be discussed in Section 6.3.

In general, the top 10 candidates could lead to over 80% accuracy at the subclass level (much improved from previous results before removing duplicates in Cilin, where it usually took the top 15 candidates to reach about 80% accuracy). At the semantic head level, the top 50 candidates could lead to over 70% accuracy for HK and TW words and to 80% or more for BJ and SG words. The accuracy, nevertheless, is also dependent on the datasets from which features were extracted, as shown in Sections 5.2 and 5.3 above and further discussed below.

6.2 Regional Variation

The various Chinese speech communities might differ not only in the lexical items they use, but also in the way they use the lexical items in common. The demand on cross-cultural knowledge thus poses a challenge for building a Pan-Chinese lexical resource manually. Cilin, for instance, is quite biased in language use in Mainland China, and it requires experts with knowledge of a wide variety of Chinese terms to be able to manually classify lexical items specific to other Chinese speech communities. It is therefore even more important to devise robust ways for automatic classification of words from various regions.

The data heterogeneity effect is quite different for the classification of SG words and TW words, but apparently not very significant for HK words. Beyond the top 5 to 10 candidates, features extracted from the SG subcorpus for Cilin words seem to have an advantage. This suggests that although the SG subcorpus shares those words in Cilin, the context in which they are used might be slightly different from their use in Mainland China. Thus extracting their contextual features from the SG subcorpus might better reflect their usage and make them more comparable with the target words from SG. For the TW words, on the contrary, features for Cilin words extracted from the BJ subcorpus always have an advantage over those extracted from the TW subcorpus. As Kwong and Tsou (2006) observed, Beijing and Taipei data share the least number of lexical items among the four regions under investigation. Words in Cilin therefore might not have the appropriate contextual feature vectors extracted from the TW subcorpus.

6.3 Analysis for Individual Words

In order to study the actual effect of various experimental conditions on the classification of individual target words, we also worked out the change in the ranking (Δr) of the correct category for each target word. A negative Δr thus corre-

sponds to an improvement in the classification as the new ranking of the correct category is smaller (earlier) than the old one. Table 3 shows some examples with improvement in this regard. The Rank column refers to the rank of the correct category in Sub-{bj,hk,tw,sg}-bj, $\Delta r(D)$ is the change after duplicates were removed from Cilin, and $\Delta r(H)$ is the change from Sub-*x*-*x* conditions.

No.	Word (Region)	Rank	$\Delta r(D)$	$\Delta r(H)$
1	信息化 (BJ)	1	-6	-
2	再就業 (BJ)	3	-15	-
3	下崗 (BJ)	1	-1	-
4	抗旱 (BJ)	1	-2	-
5	質檢 (BJ)	3	-2	-
6	大市 (HK)	4	-4	-6
7	入市 (HK)	2	-3	-6
8	國企股 (HK)	1	-6	-1
9	息口 (HK)	1	-8	-5
10	沽售 (HK)	2	-4	-5
11	新元 (SG)	1	-3	-4
12	馬股 (SG)	1	-12	-1
13	閉市價 (SG)	1	-3	-1
14	附加股 (SG)	2	-2	-4
15	容積率 (SG)	2	-7	1
16	投信 (TW)	3	-1	-13
17	成長率 (TW)	2	-3	-7
18	金控 (TW)	1	0	-3
19	買超 (TW)	6	-3	-20
20	行庫 (TW)	3	-6	-5

Table 3 Ranking Change for Individual Words[2]

Take the example of the BJ target word 信息化 (*xin4xi1hua4*, informationize). Before duplicated words were removed from Cilin, the most appropriate subclass (Ih:Change) ranked 7th in the automatic classification. Upon the removal of duplicated words, subclass Ih ranked first in the results. The words shared by other top ranking subclasses (e.g. Je:influence, Da:condition, etc.) such as 加 (*jia1*, increase), 推 (*tui1*, push), 提高 (*ti2gao1*, raise), etc., may have skewed the similarity comparison by introducing many common co-occurring words which are not particularly characteristic of any subclass.

For the TW target word 投信 (*tou2xin4*, investment trust), the appropriate subclass (Dm:organization) soared from the 16th to the 3rd when features were extracted for the Cilin words from the BJ subcorpus instead of the TW subcorpus. It was observed that both vectors have a large part in common, but the one extracted from TW subcorpus contained many more spurious features which might not be characteristic of the subclass, thus affecting the similarity score.

It is also apparent that region-specific but common concepts like 寫字樓 (*xie3zi4lou2*, office), 組屋 (*zu3wu1*, apartment), and 私宅 (*si1zhai2*, private residence), etc., are more adversely affected when features for Cilin words were extracted from the BJ subcorpus instead of the respective target subcorpora, while other more core financial concepts could often take advantage of the former. Thus it appears that the domain and concept specificity could also affect the effectiveness of the method.

6.4 Future Directions

There is room to improve the results at both the subclass and semantic head level. More qualitative analysis is needed for the data heterogeneity effect. The category size, and as pointed out above, the domain and concept specificity are also worth further investigation. The latter will thus involve the classification of words from other special domains like sports, as well as those from the general domain.

One problem we need to address in the next step is the class imbalance problem as Cilin categories could differ considerably in size, which will affect the number of features and subsequent classification. For this we plan to try the *k nearest neighbours* approach. In addition, the features might need to be constrained, as simple co-occurrence might be too coarse for distinguishing the subtle characteristics among Cilin categories.

7 Conclusion

We have worked on extending a Chinese thesaurus with words distinctly used in various Chinese communities. Classification results have improved as duplicated words in Cilin were removed. In view of the demand on cross-cultural knowledge for building a Pan-Chinese lexical resource manually, it is particularly important to devise robust ways for automatic acquisition of such a resource. Automatic classification of words with respect to an existing classificatory structure with proper datasets for feature extraction should be a prominent direction in this re-

[2] English gloss: 1-informationize, 2-re-employed, 3-unemployed, 4-resist drought, 5-quality check, 6-general trend of stock market, 7-buy in stocks, 8-H stock, 9-interest rate, 10-sell (stocks), 11-Singaporean dollar, 12-Malaysian stocks, 13-closing price, 14-rights issue, 15-holding space rate, 16-investment trust, 17-growth rate, 18-financial holdings, 19-over-bought, 20-bank.

gard. Further investigation is needed to better understand the interaction among data heterogeneity, category size, feature selection, and the domain and concept specificity of the words.

Acknowledgements

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 1317/03H). The authors would like to thank Jingbo Zhu for useful discussions on an earlier draft of this paper, and the anonymous reviewers for their comments on the submission.

References

Caraballo, S.A. (1999) Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, Maryland, USA, pp.120-126.

Ciaramita, M. (2002) Boosting automatic lexical acquisition with morphological information. In *Proceedings of the ACL'02 Workshop on Unsupervised Lexical Acquisition*, Philadelphia, USA, pp.17-25.

Hearst, M. (1992) Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, Nantes, France, pp.539-545.

Kwong, O.Y. and Tsou, B.K. (2006) Feasibility of Enriching a Chinese Synonym Dictionary with a Synchronous Chinese Corpus. In T. Salakoski, F. Ginter, S. Pyysalo and T. Pahikkala (Eds.), *Advances in Natural Language Processing: Proceedings of FinTAL 2006.* Lecture Notes in Artificial Intelligence, Vol.4139, pp.322-332, Springer-Verlag.

Kwong, O.Y. and Tsou, B.K. (2007) Extending a Thesaurus in the Pan-Chinese Context. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-2007)*, Prague, pp.325-333.

Lin, D. (1998) Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, Montreal, Canada, pp.768-774.

Mei *et al.* 梅家駒、竺一鳴、高蘊琦、殷鴻翔 (1984) 《同義詞詞林》 (*Tongyici Cilin*). 商務印書館 (Commercial Press) / 上海辭書出版社.

Pekar, V. (2004) Linguistic Preprocessing for Distributional Classification of Words. In *Proceedings of the COLING2004 Workshop on Enhancing and Using Electronic Dictionaries*, Geneva.

Riloff, E. and Shepherd, J. (1997) A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, Rhode Island, pp.117-124.

Thelen, M. and Riloff, E. (2002) A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, USA.

Tokunaga, T., Fujii, A., Iwayama, M., Sakurai, N. and Tanaka, H. (1997) Extending a thesaurus by classifying words. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, pp.16-21.

Tseng, H. (2003) Semantic Classification of Chinese Unknown Words. In the *Proceedings of the ACL-2003 Student Research Workshop, Companion Volume to the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan.

Tsou, B.K. and Kwong, O.Y. (2006) Toward a Pan-Chinese Thesaurus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Tsou, B.K. and Lai, T.B.Y. 鄒嘉彥、黎邦洋 (2003) 漢語共時語料庫與信息開發. In B. Xu, M. Sun and G. Jin 徐波、孫茂松、靳光瑾 (Eds.), 《中文信息處理若干重要問題》 (*Issues in Chinese Language Processing*). 北京：科學出版社, pp.147-165

You, J-M. and Chen, K-J. (2006) Improving Context Vector Models by Feature Clustering for Automatic Thesaurus Construction. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, COLING-ACL 2006, Sydney, Australia, pp.1-8.