Extending a Thesaurus in the Pan-Chinese Context

Oi Yee Kwong and Benjamin K. Tsou
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 address a unique problem in Chinese language processing and report on our study on extending a Chinese thesaurus with *region-specific* words, mostly from the financial domain, from various Chinese speech communities. With the larger goal of automatically constructing a Pan-Chinese lexical resource, this work aims at taking an existing semantic classificatory structure as leverage and incorporating new words into it. In particular, it is important to see if the classification could accommodate new words from heterogeneous data sources, and whether simple similarity measures and clustering methods could cope with such variation. We use the cosine function for similarity and test it on automatically classifying 120 target words from four regions, using different datasets for the extraction of feature vectors. The automatic classification results were evaluated against human judgement, and the performance was encouraging, with accuracy reaching over 85% in some cases. Thus while human judgement is not straightforward and it is difficult to create a Pan-Chinese lexicon manually, it is observed that combining simple clustering methods with the appropriate data sources appears to be a promising approach toward its automatic construction.

1 Introduction

Large-scale semantic lexicons are important resources for many natural language processing (NLP) tasks. For a significant world language such as Chinese, it is especially critical to capture the substantial *regional variation* as an important part of the lexical knowledge, which will be useful for many NLP applications, including natural language understanding, information retrieval, and machine translation. Existing Chinese lexical resources, however, are often based on language use in one particular region and thus lack the desired comprehensiveness.

Toward this end, Tsou and Kwong (2006) proposed a comprehensive Pan-Chinese lexical resource, based on a large and unique synchronous Chinese corpus as an authentic source for lexical acquisition and analysis across various Chinese speech communities. To allow maximum *versatility* and *portability*, it is expected to document the core and universal substances of the language on the one hand, and also the more subtle variations found in different communities on the other. Different Chinese speech communities might share lexical items in the same form but with different meanings. For instance, the word 居屋 refers to general housing in Mainland China but specifically to housing under the Home Ownership Scheme in Hong Kong; and while the word 住房 is similar to 居屋 to mean general housing in Mainland China, it is rarely seen in the Hong Kong context.

Hence, the current study aims at taking an existing Chinese thesaurus, namely the *Tongyici Cilin* 同義詞詞林, as leverage and extending it with lexical items specific to individual Chinese speech communities. In particular, the feasibility depends on the following issues: (1) Can lexical items from various Chinese speech communities, that is, from such heterogeneous sources, be classified as effectively with methods shown to work for clustering

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 325–333, Prague, June 2007.

closely related words from presumably the same, or homogenous, source? (2) Could existing semantic classificatory structures accommodate concepts and expressions specific to individual Chinese speech communities?

Measuring similarity will make sense only if the feature vectors of the two words under comparison are directly comparable. There is usually no problem if both words and their contextual features are from the same data source. Since Tongyici Cilin (or simply Cilin hereafter) is based on the vocabulary used in Mainland China, it is not clear how often these words will be found in data from other places, and even if they are found, how well the feature vectors extracted could reflect the expected usage or sense. Our hypothesis is that it will be more effective to classify new words from Mainland China with respect to Cilin categories, than to do the same on new words from regions outside Mainland China. Furthermore, if this hypothesis holds, one would need to consider separate mechanisms to cluster heterogeneous region-specific words in the Pan-Chinese context.

Thus in the current study we sampled 30 target words specific to each of Beijing, Hong Kong, Singapore, and Taipei, from the financial domain; and used the cosine similarity function to classify them into one or more of the semantic categories in Cilin. The automatic classification results were compared with a simple baseline method, against human judgement as the gold standard. In general, an accuracy of up to 85% could be reached with the top 15 candidates considered. It turns out that our hypothesis is supported by the Taipei test data, whereas the data heterogeneity effect is less obvious in Hong Kong and Singapore test data, though the effect on individual test items varies.

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

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; Thelen and Riloff, 2002). For instance, Lin (1998) used dependency relation as word features to compute word similarities from large corpora, and compared the thesaurus created in such a way with WordNet and Roget classes. Caraballo (1999) selected head nouns from conjunctions and appositives in noun phrases, and used the cosine similarity measure with a bottom-up clustering technique to construct a noun hierarchy from text. Curran and Moens (2002) explored a new similarity measure for automatic thesaurus extraction which better compromises with the speed/performance tradeoff. You and Chen (2006) used a feature clustering method to create a thesaurus from a Chinese newspaper corpus.

Another line of research, which is more closely related with 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.

The current work attempts to classify new words with an existing thesaural classificatory structure. However, the usual practice in past studies is to test with a portion of data from the thesaurus itself and evaluate the results against the original classification of those words. This study is thus different in the following ways: (1) The test data (i.e. the target words to be classified) were not taken from the thesaurus, but extracted from corpora and these words were unknown to the thesaurus. (2) The target words were not limited to nouns. (3) Automatic classification results were compared with a baseline method and with the manual judgement of several linguistics students constituting the gold standard. (4) In view of the heterogeneous nature of the Pan-Chinese context, we experimented with extracting feature vectors from different datasets.

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. The Appendix shows some example subclasses. In the following discussion, we will mainly refer to the subclass level and semantic head level.

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, Shanghai, 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 2006, the corpus has already accumulated over 200 million character tokens which, upon automatic word segmentation and manual verification, amount to over 1.2 million word types.

For the present study, we made use of the subcorpora collected over the 9-year period 1995-2004 from Beijing (BJ), Hong Kong (HK), Singapore (SG), and Taipei (TW). In particular, we made use of the *financial news* sections in these subcorpora, from which we extracted feature vectors for comparing similarity between a given target word and a thesaurus class, which is further explained in Section 4.3. Table 1 shows the sizes of the subcorpora.

3.3 Test Data

Instead of using a portion of Cilin as the test data, we extracted unique lexical items from the various subcorpora above, and classified them with respect to the Cilin classification.

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 from each of these unique item lists for testing. Classification was based on their similarity with each of the Cilin subclasses, compared by the cosine measure, as discussed in Section 4.3.

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

Table 1 Sizes of Individual Subcorpora

4 Experiments

4.1 Human Judgement

Three undergraduate linguistics students and one research student on computational linguistics from the City University of Hong Kong were asked to do the task. The undergraduate students were raised in Hong Kong and the research student in Mainland China. They were asked to assign what they consider the most appropriate Cilin category (up to the semantic head level, i.e. third level in the

Cilin structure) to each of the 120 target words. The inter-annotator agreement was measured by the *Kappa* statistic (Siegel and Castellan, 1988), at both the subclass and semantic head levels. Results on the human judgement are discussed in Section 5.1.

4.2 Creating Gold Standard

The "gold standard" was set at both the subclass level and semantic head level. For each level, we formed a "strict" standard for which we considered all categories assigned by at least two judges to a word; and a "loose" standard for which we considered all categories assigned by one or more judges. For evaluating the automatic classification in this study, however, we only experimented with the loose standard at the subclass level.

4.3 Automatic Classification

Each target word was automatically classified with respect to the Cilin subclasses based on the similarity between the target word and each subclass.

We compute the similarity by the cosine between the two corresponding feature vectors. The feature vector of a given target word contains *all its co-occurring content words* in the 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 subclass is based on the union of the features (i.e. co-occurring words in the corpus) from all individual members in the subclass.

The cosine of two feature vectors is computed as

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

In view of the difference in the feature space of a target word and a whole class of words, and thus the potential difference in the number of occurrence of individual features, we experimented with two versions of the cosine measurement, namely binary vectors and real-valued vectors.

In addition, as mentioned in previous sections, we also experimented with the following conditions: whether feature vectors for the Cilin subclasses were extracted from the subcorpus where a given target word originates, or from the Beijing subcorpus which is assumed to be representative of language use in Mainland China. All automatic classification results were evaluated against the gold standard based on human judgement.

4.4 Baseline

To evaluate the effectiveness of the automatic classification, we adopted a simple baseline measure by ranking the 94 subclasses in descending order of the number of words they cover. In other words, assuming the bigger the subclass size, the more likely it covers a new term, thus we compared the top-ranking subclasses with the classifications obtained from the automatic method using the cosine measure.

5 Results and Discussion

5.1 Response from Human Judges

All human judges reported difficulties in various degrees in assigning Cilin categories to the target words. The major problem comes from the regional specificity and thus the unfamiliarity of the judges with the respective lexical items and contexts. For instance, students grown up in Hong Kong were most familiar with the Hong Kong data, and slightly less so with the Beijing data, but more often had the least ideas for the Taipei and Singapore data. The research student from Mainland China had no problem with Beijing data and the lexical items in Cilin, but had a hard time figuring out the meaning for words from Hong Kong, Taipei and Singapore. For example, all judges reported problem with the term 自撮, one of the target words from Singapore referring to 自撮股市 (CLOB in the Singaporean stock market), which is really specific to Singapore.

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 acquisition of such a resource.

Notwithstanding the difficulty, the inter-annotator agreement was quite satisfactory. At the subclass level, we found K=0.6870. At the semantic head level, we found K=0.5971. Both figures are statistically significant.

5.2 Gold Standard

As mentioned, we set up a loose standard and a strict standard at both the subclass and semantic head level. In general, the judges managed to reach some consensus in all cases, except for two words from Singapore. For these two cases, we considered all categories assigned by any of the judges for both standards.

The gold standards were verified by the authors. Although in several cases the judges did not reach complete agreement with one another, we found that their decisions reflected various possible perspectives to classify a given word with respect to the Cilin classification; and the judges' assignments, albeit varied, were nevertheless reasonable in one way or another.

5.3 Evaluating Automatic Classification

In the following discussion, we will refer to the various testing conditions for each group of target words with labels in the form of Cos-<Vector Type>-<Target Words>-<Cilin Feature Source>. Thus the label Cos-Bin-hk-hk means testing on Hong Kong target words with binary vectors and extracting features for the Cilin words from the Hong Kong subcorpus; and the label Cos-RV-sg-bj means testing on Singapore target words with real-valued vectors and extracting features for the Cilin words from the Beijing subcorpus. For each target word, we evaluated the automatic classification (and the baseline ranking) by matching the human decisions with the top N candidates. If any of the categories suggested by the human judges is covered, the automatic classification is considered accurate. The results are shown in Figure 1 for test data from individual regions.

Overall speaking, the results are very encouraging, especially in view of the number of categories (over 90) we have at the subclass level. An accuracy of 80% or more is obtained in general if the top 15 candidates were considered, which is much higher than the baseline result in all cases. Table 2 shows some examples with appropriate classification within the Top 3 candidates. The two-letter codes in the "Top 3" column in Table 2 refer to the subclass labels, and the code in bold is the one matching human judgement.

In terms of the difference between binary vectors and real-valued vectors in the similarity measurement, the latter almost always gave better results. This was not surprising as we expected by using real-valued vectors we could be less affected by the potential huge difference in the feature space and the number of occurrence of the features for a Cilin subclass and a target word.

As for extracting features for Cilin subclasses from the Beijing subcorpus or other subcorpora, the difference is more obvious for the Singapore and Taipei target words. We will discuss the results for each group of target words in detail below.

5.4 Performance on Individual Sources

Target words from Beijing were expected to have a relatively higher accuracy because they are homogenous with the Cilin content. It turned out, however, the accuracy only reached 73% with top 15 candidates and 83% with top 20 candidates even under the Cos-RV-bj-bj condition. Words like 非典 (SARS), 節水 (save water), 產業化 (industrialize / industrialization), 合格率 (passing rate) and 傳銷 (multi-level marketing) could not be successfully classified.

Results were surprisingly good for target words from the Hong Kong subcorpus. Under the Cos-RV-hk-hk condition, the accuracy was 87% with top 15 candidates and even over 95% with top 20 candidates considered. Apart from this high accuracy, another unexpected observation is the lack of significant difference between Cos-RV-hk-hk and Cos-RV-hk-bj. One possible reason is that the relatively larger size of the Hong Kong subcorpus might have allowed enough features to be extracted even for the Cilin words. Nevertheless, the similar results from the two conditions might also suggest that the context in which Cilin words are used might be relatively similar in the Hong Kong subcorpus and the Beijing subcorpus, as compared with other communities.

Similar trends were observed from the Singapore target words. Looking at Cos-RV-sg-sg and Cos-RV-sg-bj, it appears that extracting feature vectors for the Cilin words from the Singapore subcorpus leads to better performance than extracting them from the Beijing subcorpus. It suggests that although the Singapore 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 Singapore subcorpus might better reflect their usage and makes it more comparable with the unique target words from Singapore. Such possible difference in contextual features with shared lexical items between different Chinese speech communities would require further investigation, and will form part of our future work as discussed below. Despite the above observation from the accuracy figures, the actual effect, however, seems to vary on individual lexical items. Table 3 shows some examples of target words which received similar (with white cells) and very different (with shaded cells) classification respectively under the two conditions. It appears that the region-specific but common concepts like 寫字樓 (office), 組屋 (apartment), 私宅 (private residence), which relate to building or housing, were affected most.

Taipei data, on the contrary, seems to be more affected by the different testing conditions. Cos-Bin-tw-bj and Cos-RV-tw-bj produced similar results, and both conditions showed better results than Cos-RV-tw-tw. This supports our hypothesis that the effect of data heterogeneity is so apparent that it is much harder to classify target words unique to Taipei with respect to the Cilin categories. In addition, as Kwong and Tsou (2006) observed, Beijing and Taipei data share the least number of lexical items, among the four regions under investigation. Hence, words in Cilin might not have the appropriate contextual feature vectors extracted from the Taipei subcorpus.

The different results for individual regions might be partly due to the endocentric and exocentric nature of influence in lexical innovation (e.g. Tsou, 2001) especially with respect to the financial domain and the history of capitalism in individual regions. This factor is worth further investigation.

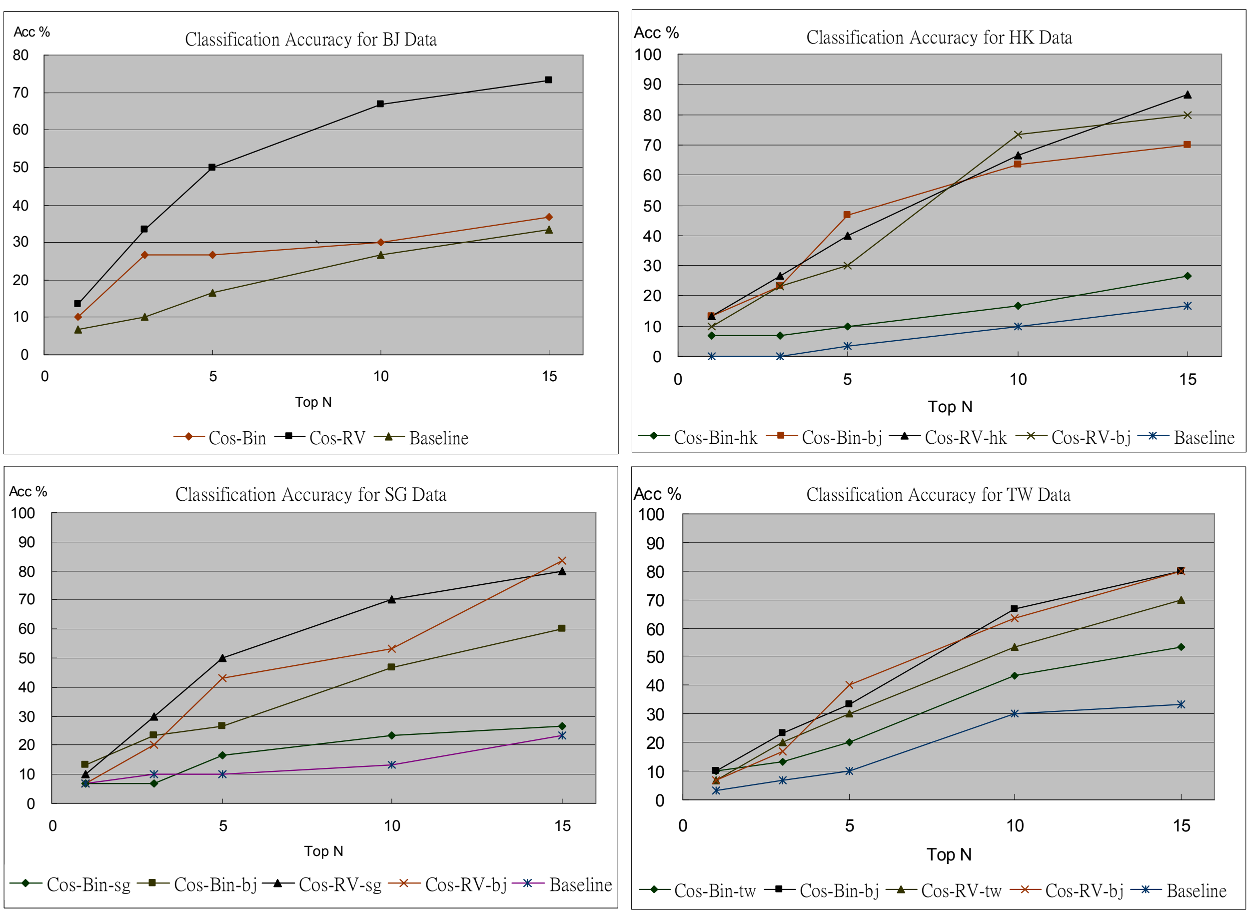

Figure 1 Classification Results with Top N Candidates

No.	Region	Word	Top 3
1	BJ	退耕還林	**Di** Gb Df
2	BJ	面料	**Bq** Ae Hd
3	BJ	煤礦	**Bm** Hi Hd
4	BJ	抓好	**Hj** Di Hd
5	BJ	下崗	Aa **If** Ae
6	HK	銷情	**Da** Cb Bi
7	HK	寬頻	**Bb** Jc Hi
8	HK	紅籌	**Dj** Da Hi
9	HK	息率	Bi **Dj** Dn
10	HK	地產股	Bi **Dj** Gb
11	SG	財年	**Ca** Dm Hi
12	SG	賣空	**Ig** He Dj
13	SG	獻議	Dm Dj **Hi**
14	SG	脫售	Dm Dj **He**
15	SG	准將	Hi Hg **Af**
16	TW	金控	**Dm** Hd Hi
17	TW	個股	Jb Dn **Dj**
18	TW	房市	Ja Ca **He**
19	TW	現金卡	Hf **Dj** Dm
20	TW	存底	**Dj** Ed Ca

Table 2 Examples of Correct Classification (Top 3)[1]

5.5 General Discussions and Future Work

As mentioned in a previous section, the test data in this study were not taken from the thesaurus itself, but were unknown words to the thesaurus. They were extracted from corpora, and were not limited to nouns. We found in this study that the simple cosine measure, which used to be applied for clustering contextually similar words from homogenous sources, performs quite well in general for classifying these unseen words with respect to the Cilin subclasses. The automatic classification results were compared with the manual judgement of several linguistics students. In addition to providing a gold standard for evaluating the automatic classification results in this study, the human judgement on the one hand proves that the Cilin classificatory structure could accommodate region-specific lexical items; but on the other hand also suggests how difficult it would be to construct such a Pan-Chinese lexicon manually as rich cultural and linguistic knowledge would be required. Moreover, we started with Cilin as the established semantic classification and attempted to classify words specific to Beijing, Hong Kong, Singapore, and Taipei respectively. The heterogeneity of sources did not seem to hamper the similarity measure on the whole, provided appropriate datasets are used for feature extraction, although the actual effect seemed to vary on individual lexical items.

No.	Source	Word	Ranking of 1st appropriate class	
			Cos-RV-hk-hk, etc.	Cos-RV-hk-bj, etc.
1	HK	銷情	1	1
2	HK	寬頻	1	1
3	HK	紅籌	1	1
4	HK	加推	2	10
5	HK	低位	19	5
6	HK	寫字樓	13	30
7	SG	財政年	2	2
8	SG	賣空	2	1
9	SG	附加股	5	4
10	SG	組屋	1	12
11	SG	容積率	1	9
12	SG	私宅	8	26
13	TW	存底	1	1
14	TW	個股	4	3
15	TW	金控	5	1
16	TW	投信	18	4
17	TW	成長率	12	5
18	TW	現金卡	8	2

Table 3 Different Impact on Individual Items[2]

Despite the encouraging results with the top 15 candidates in the current study, it is desirable to improve the accuracy for the system to be useful in

[1] English gloss: 1-restoring agricultural lands for afforestation, 2-material, 3-coal mine, 4-to seize (an opportunity), 5-unemployed, 6-sales performance, 7-broadband, 8-red chip, 9-interest rate, 10-property stocks, 11-financial year, 12-sell short, 13-proposal, 14-sell, 15-brigadier general, 16-financial holdings, 17-individual stocks, 18-property market, 19-cash card, 20-stub.

[2] English gloss: 1-sales performance, 2-broadband, 3-red chip, 4-add (supply to market), 5-low level, 6-office, 7-financial year, 8-sell short, 9-rights issue, 10-apartment, 11-holding space rate, 12-private residence, 13-stub, 14-individual stocks, 15-financial holdings, 16-investment trust, 17-growth rate, 18-cash card.

practice. Hence our next step is to expand the test data size and to explore alternative methods such as using a nearest neighbour approach. In addition, we plan to further the investigation in the following directions. First, we will experiment with the automatic classification at the Cilin semantic head level, which is much more fine-grained than the subclasses. The fine-grainedness might make the task more difficult, but at the same time the more specialized grouping might pose less ambiguity for classification. Second, we will further experiment with classifying words from other special domains like sports, as well as the general domain. Third, we will study the classification in terms of the part-of-speech of the target words, and their respective requirements on the kinds of features which give best classification performance.

The current study only dealt with presumably Modern Standard Chinese in different communities, and it could potentially be expanded to handle various dialects within a common resource, eventually benefiting speech lexicons and applications at large.

6 Conclusion

In this paper, we have reported our study on a unique problem in Chinese language processing, namely extending a Chinese thesaurus with new words from various Chinese speech communities, including Beijing, Hong Kong, Singapore and Taipei. The critical issues include whether the existing classificatory structure could accommodate concepts and expressions specific to various Chinese speech communities, and whether the difference in textual sources might pose difficulty in using conventional similarity measures for the automatic classification. Our experiments, using the cosine function to measure similarity and testing with various sources for extracting contextual vectors, suggest that the classification performance might depend on the compatibility between the words in the thesaurus and the sources from which the target words are drawn. Evaluated against human judgement, an accuracy of over 85% was reached in some cases, which were much higher than the baseline and were very encouraging in general. While human judgement is not straightforward and it is difficult to create a Pan-Chinese lexicon manually, combining simple classification methods with the appropriate data sources seems to be a promising approach toward its automatic construction.

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).

Appendix

The following table shows some examples of the Cilin subclasses:

Class	Subclasses
A 人 (Human)	Aa … Ae 職業 (Occupation) Af 身份 (Identity) … An
B 物 (Things)	Ba … Bb 擬狀物 (Shape) … Bi 動物 (Animal)… Bm 材料 (Material)…Bq 衣物 (Clothing) … Br
C 時間與空間 (Time and Space)	Ca 時間 (Time) Cb 空間 (Space)
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)
E 特徵 (Characteristics)	Ea … Ed 性質 (Property)… Ef
F 動作 (Action)	Fa … Fd
G 心理活動 (Psychological activities)	Ga … Gb 心理活動 (Psychological activities)… Gc
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)
I 現象與狀態 (Phenomenon and state)	Ia … If 境遇 (Circumstance) Ig 始末 (Process)… Ih
J 關聯 (Association)	Ja 聯繫 (Liaison) Jb 異同 (Similarity and Difference) Jc 配合 (Matching) … Je
K 助語 (Auxiliary words)	Ka … Kf
L 敬語 (Respectful expressions)	

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.

Curran, J.R. and Moens, M. (2002) Improvements in Automatic Thesaurus Extraction. In *Proceedings of the ACL'02 Workshop on Unsupervised Lexical Acquisition*, Philadelphia, USA, pp.59-66.

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.

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 (Commerical 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.

Siegel, S. and Castellan, N.J. (1988) *Nonparametric Statistics for the Behavioral Sciences (2nd Ed.)*. McGraw-Hill.

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. (2001) Language Contact and Lexical Innovation. In M. Lackner, I. Amelung and J. Kurtz (Eds.), *New Terms for New Ideas: Western Knowledge and Lexical Change in Late Imperial China*. Berlin: Brill.

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.