Learning Semantic Lexicons Using Graph Mutual Reinforcement Based Bootstrapping

ZHANG Qi[1] QIU Xi-Peng[1] HUANG Xuan-Jing[1] WU Li-De[1]

Abstract This paper presents a method to learn semantic lexicons using a new bootstrapping method based on graph mutual reinforcement (GMR). The approach uses only unlabeled data and a few seed words to learn new words for each semantic category. Different from other bootstrapping methods, we use GMR-based bootstrapping to sort the candidate words and patterns. Experimental results show that the GMR-based bootstrapping approach outperforms the existing algorithms both in in-domain data and out-domain data. Furthermore, it shows that the result depends on not only the size of the corpus but also the quality.

Key words Semantic lexicon, bootstrapping, graph mutual reinforcement (GMR)

In recent years, bootstrapping methods[1−3] have received considerable attention in many application fields, and semantic lexicons[4−6] have proved useful for many natural language processing tasks. Although supervised methods usually can achieve better results than those by semi-supervised and unsupervised methods, they are strongly constrained by the number of labeled data. Usually, bootstrapping methods can use both a small size of labeled data and a large amount of unlabeled samples without extra cost to obtain better result.

Semantic lexicons have proved to be useful for many natural language processing tasks, including question answering[7−8], information extraction[9] and so on. Learning semantic lexicons is a task to automatically acquire words with semantic classes (e.g., "gun" is a WEAPON). In recent years, several algorithms have been proposed to automatically learn semantic lexicons using supervised, semi-supervised, and unsupervised methods[10−13]. As unsupervised methods dispense with manually labeled training data, they have attracted increased attentions[3, 14−18].

In this paper, we introduce a weakly supervised learning method, graph mutual reinforcement (GMR) based bootstrapping, called GMR-bootstrapping, to learn semantic lexicons. Like other bootstrapping methods, it begins with unlabeled corpus and a few seed words. Then, it is iterated to learn lexicons. From analyzing the procedure of Basilisk[3], we found that the patterns which contain a large amount of extractions always obtained with very low scores at the beginning. In order to partially overcome this problem, we incorporated GMR to weight candidate words and extraction patterns. The similar idea was also used by Hassan[16] to select the informative patterns for extracting information. Normally, if the extractions of a pattern belong to several different categories, the extraction accuracy is low. To better use this information, we also enhanced the GMR-bootstrapping by adding the uncertainty of a pattern into scoring functions to learn multiple categories simultaneously.

Evaluation on MUC4 corpus[19] shows that incorporating GMR to weight the candidate words and extraction patterns enables substantial performance gains in extracting BUILDING, EVENT, HUMAN, LOCATION, TIME, and WEAPON lexicons. Our experimental results showed that adding patterns′ uncertainty into scoring functions improve the performance also. From the experimental results, we also observed that the quality of lexicons of automobile manufacture names and automobile parts extracted by GMR-bootstrapping from Chinese corpus (detailed in Section 2) was better than the quality of lexicons extracted by Basilisk. The reminder of the paper is organized as follows: In Section 1, we introduce our bootstrapping structure and scoring functions. In Section 2, experiments are given to show the improvements. Section 3 discusses the related works. Section 4 concludes the paper.

Received June 21, 2007; in revised form January 11, 2008
Supported by National Natural Science Foundation of China (60673038, 60503070)
1. Department of Computer Science and Technology, Fudan University, Shanghai 200433, P. R. China
DOI: 10.3724/SP.J.1004.2008.01257

1 GMR-bootstrapping

GMR-bootstrapping[20] is a weakly supervised learning method. Like other bootstrapping methods, the inputs of GMR-bootstrapping are a large amount of unlabeled data and a few manually selected seed words for each semantic category. The GMR-bootstrapping begins with extracting a number of the extraction patterns that can match the seed words. Then, a number of nouns extracted by these patterns become candidates for the lexicon. Then, a bipartite graph is built, which represents the matching relation between patterns and candidate words. Next, GMR scoring is used to iteratively assign correctness weights of patterns and candidate words. The five best candidate words are added to the lexicon. Then, process starts over again. In this section, we describe details of the GMR-bootstrapping algorithm.

1.1 Pattern formats

In order to find new lexicon entries, extraction patterns are used to provide the contextual evidence that a word belongs to certain semantic class. There are two commonly used patterns, Syntactic Pattern (SP) and Context Pattern. SPs are used by many other bootstrapping methods[2−3, 21]. We followed the method proposed by Riloff[22], which used the AutoSlog system to represent extraction patterns. AutoSlog′s extraction patterns represent linguistic expressions that extract the head noun of a noun phrase in one of three syntactic roles: subject, direct object, or prepositional phrase object. Different from the syntactic pattern, context patterns use words only, where it excludes syntactic roles. In our implementations, we used words before or after current words as templates in MUC4 corpus.

1.2 GMR scoring

GMR scoring is used to iteratively assign the scores of patterns and candidate words. We assume that patterns that match many words from the same category tend to be important. Similarly, words matched by many patterns that belong to a same category tend to be correct[23].

Each pattern p in P is associated with a weight $sp(p)$ denoting its correctness. Each candidate word w in W has

a weight $sw(w)$, which expresses the correctness of the word. The weights are calculated iteratively through (1) to (7) as

$$F^{(i)}(p) = \sum_{u \in W(p)} sw(u)l \tag{1}$$

$$sp^{(i)}(p) = \frac{F^{(i)}(p) \cdot \ln F^{(i)}(p)}{|W(p)|} \tag{2}$$

$$sw^{(i)}(w) = \frac{\sum p \in P(w) \ln (F^{(i)}(p) + 1)}{|P(w)|} \tag{3}$$

where $sw(u)$ is initialized to 1 if $u \in$ Semantic Lexicons, and 0 otherwise, $W(p)$ is the set of words matched by p, $P(w)$ is the set of patterns matching w, $sp^{(i)}(p)$ is the correct weight of the pattern p in iteration i, and $sw^{(i)}(w)$ is the correct weight of the word w in iteration i.

At the end of each iteration, the normalization factors, $SP^{(i)}$ and $SW^{(i)}$, are calculated as

$$SP^{(i)} = \sum_{w=1}^{|W|} \sum_{v=1}^{P(w)} sp^{(i)}(v) \tag{4}$$

$$SW^{(i)} = \sum_{p=1}^{|P|} \sum_{u=1}^{W(p)} sw^{(i)}(u) \tag{5}$$

Then, $sp^{(i)}(p)$ and $sw^{(i)}(w)$ are normalized by

$$sp^{(i)}(p) = \frac{sp^{(i)}(p)}{SP^{(i)}} \tag{6}$$

$$sw^{(i)}(w) = \frac{sw^{(i)}(w)}{SW^{(i)}} \tag{7}$$

Here, (2) is similar to *RlogF*, which has been used to score patterns[3], except that F_i in *RlogF* is changed to (1). *RlogF* becomes

$$RlogF(pattern_i) = \frac{F_i}{N_i} \cdot \ln(F_i)$$

where F_i is the number of the distinct categories extracted by $pattern_i$ and N_i is the total number of the nouns extracted by $pattern_i$. Because the amount of category numbers depends on the number of seed words, which is usually small, F_i is a small value at the first 10 iterations. However, N_i may be a big value. Due to this observation and the the definition of *RlogF*, the patterns which contain a large number of extractions would obtain low scores by *RlogF* at the beginning though some of them are good. For example, pattern k "HIT BY *" is a good pattern for the WEAPON category. 82.3 % of words extracted by this pattern belong to this category. However, in the first iteration, F_k is 1, because there are only three words in the category, while N_k, the number of words extracted by pattern k, is 17. According to the equation *RlogF*, the score of this pattern is 0. It is equal to other patterns with one extraction. Equation (3) is changed from *AvgLog*, which has been used to score candidate words[3]. Through these changes, the scoring functions of patterns and candidate words are connected and can be iteratively calculated. The scores of good patterns with a large amount of extractions will be increased with iterations. Since (2) is based not only on the lexicons at this stage but also on the score of other words, the patterns like "HIT BY *" are given a bigger score than the *AvgLog*. The scores of candidate words extracted by these patterns are also improved. Consequently, the problem of *RlogF* can be partially overcomed.

1.3 Learning multiple semantic categories

From the Thelen's analysis and results of Basilisk-MACT+[3], we observe that learning multiple semantic categories can improve the results of all the categories. We also extended GMR-bootstrapping to learn multiple semantic categories simultaneously, named GMR-M-bootstrapping. Normally, if the extraction of a pattern belongs to several different categories, the pattern's correctness should be low.

We used L_p to represent the labels of the extractions of the pattern p. $H(L_p)$ is the entropy of L_p, which is calculated only in the extractions that have been labeled to a semantic category. For instance, pattern p, whose extractions are $w_1, w_2, \cdots, w_n$. We can find the labels of its extractions through semantic lexicons at this stage, $L_p = l_1, l_2, \cdots, l_n$, where

$$l_i = \begin{cases} Label_j, & \text{if } w_i \in Lexicon_j \\ \text{NULL}, & \text{otherwise} \end{cases}$$

Then, the entropy of L_p is calculated

$$H(L_p) = -\sum_{k=1}^{|\chi|} p\left(Label_k\right) \cdot \ln\left(p\left(Label_k\right)\right) \tag{8}$$

where $p(Label_k) = \dfrac{C_k}{\sum_{k=1}^{|\chi|} C_k}$, C_k denotes the number of times $Label_k$ occurs in L_p. We could define the patterns' uncertainty, $(1 - \dfrac{H(L_p)}{\log|\chi|})$, which varies from 0 when L_p is uniform to 1 when L_p contains one type of labels[24].

Therefore, (2) and (3) can be changed to

$$sp^{(i+1)}(p) = \frac{F^{(i+1)}(p) \cdot \ln F^{(i+1)}(p) \cdot (1 - \frac{H(L_p)}{\log|\chi|})}{|W(p)|} \tag{9}$$

$$sw^{(i+1)}(w) = \frac{\sum\limits_{p \in P(w)} \ln (F^{(i+1)}(p) + 1) \cdot (1 - \frac{H(L_p)}{\log(|\chi|)})}{|P(w)|} \tag{10}$$

which are modified by multiplying (2) and (3) by the uncertainty of patterns, respectively. The experiments and results using (9) and (10) are shown in Section 2.

2 Experiments

To compare the performance of GMR-bootstrapping to other weakly supervised methods, we designed several experiments, and evaluated on two corpora. The one is the MUC-4 corpus[19], which contains 1 700 texts (includes both test and training parts) in terrorism domain. All the words in the corpus are divided into nine semantic categories[3]: BUILDING, EVENT, HUMAN, LOCATION, ORGANIZATION, TIME, VEHICLE, WEAPON and OTHER. A few semantic lexicon learners have been evaluated on this corpus[2, 3, 12, 21]. Basilisk achieved the best results. We implemented the Basilisk algorithm to compare it with GMR-bootstrapping. The other is Chinese review corpus about vehicle (CRCV), collected by us. CRCV contains about 500 000 articles in around 500 automobile domain forums. All the articles are reviews about vehicle. GMR-bootstrapping and Basilisk were used to learn MANUFACTURE and PARTS categories in this corpus. An open domain Chinese corpus, which contains about 1 000 000 news articles, was also used to evaluate the stability of GMR-bootstrapping and Basilisk.

2.1 Results in MUC4 corpus

Fig. 1 shows the results provided by GMR-bootstrapping and by repeated Basilisk(R-Basilisk) with *SP*. For each category, the top 10 most frequent nouns that belong to the category were extracted as seed words in the same way in [3]. We carried out the experiment with different patterns for 200 iterations, so 1 000 words are extracted. The X axis shows the number of words extracted. The Y axis shows the number of correct words. The results show that the performance of GMR-bootstrapping is better than R-Basilisk′s in all categories. The results also show that the R-Basilisk′s performance is similar to the Basilisks′ results reported in [3] in all categories.

The next experiment aimed to test the R-Basilisk and our algorithm′s stability with different seeds′ quality. When different seeds were used, the number of correct extractions would be changed. We randomly selected 20 seed lists (each compared list contains 10 seed words) for each category and compared the results with R-basilisk. We ran both approaches with 200 iterations with all seed list. Fig. 2(see next page) shows the results. The X axis shows the six categories. The Y axis shows the average number of correct extractions. The left column (R-Basilisk average) in each category represents the results of R-basilisk, while the right one (GMR average) is the GMR-bootstrapping′s results. Two lines represent the best results in each category of R-Basilisk and GMR-bootstrapping, respectively. The seed words which are used in R-Basilisk best and GMR best are extracted according to the seeds′ frequency in the MUC4 Corpus. The top 10 frequent words were used as seeds for each category. The results show that the average score of GMR-bootstrapping outperforms R-Basilisk in all the categories. The relative improvement is more than 100 in BUILDING, EVENT, and WEAPON categories. In LOCATION and TIME categories, the improvement is around 70 %. This indicates that GMR-bootstrapping is more reliable than R-Basilisk. Fig. 2 shows that the best result of GMR-bootstrapping is better than R-Basilisk′s in all the categories.

Then, we evaluated GMR-M-bootstrapping to learn multiple semantic categories simultaneously. The detailed results are shown in [20]. Experiments show that GMR-M-bootstrapping′s results are better than GMR-bootstrapping′s in all the categories. Those results indicate

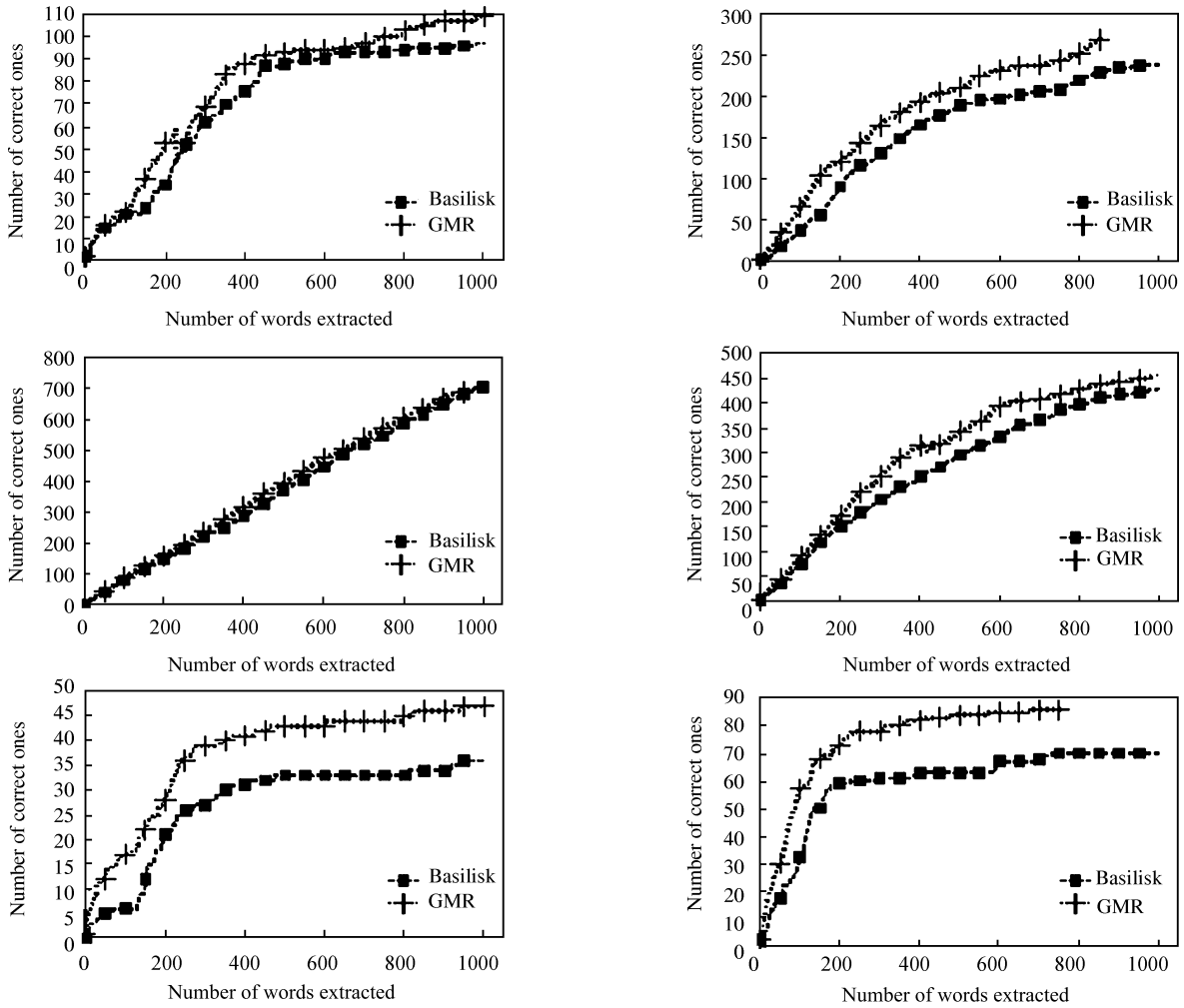


Fig. 1 GMR-bootstrapping vs. Repeated Basilisk (R-Basilisk)

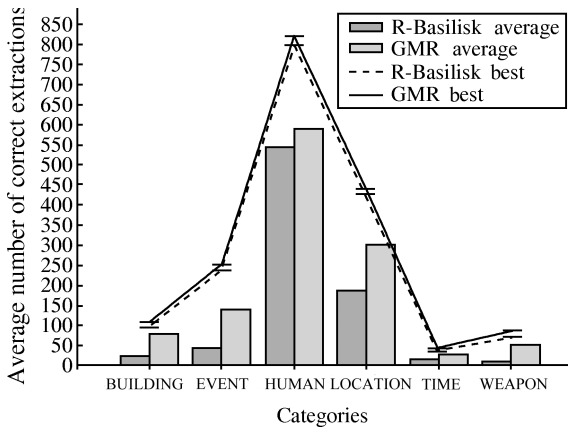


Fig. 2 GMR-bootstrapping and R-Basilisk′s best result and the average number of correct extractions.

that our method can improve the results. Pattern′s uncertainty can also benefit the final results.

2.2 Results in CRCV corpus

We also compared the results of GMR-bootstrapping with R-Basilisk in CRCV Corpus (detailed in the beginning of the Section 2. We ran both of approaches with 200 iterations to learn MANUFACTURE and PARTS categories. A total of 10 seed words were used for each category. The results in Table 1 show that GMR-M-bootstrapping′s performance is better than GMR-bootstrapping′s and GMR-bootstrapping′s performance is better than R-basilisk′s in both categories. The trend is the same as the results in MUC-4 corpus.

Table 1 GMR-bootstrapping vs. Basilisk in Chinese CRCV corpus

Category	Total words	GMR	GMR-M	R-Basilisk
MANUFACTURE	100	47	**61**	44
	200	67	**82**	61
	944	107	**120**	92
PARTS	100	59	**72**	49
	200	110	**126**	59
	1000	282	**289**	192

In order to evaluate the corpus′s impact, we randomly selected 10 %, 20 %, 40 %, and 80 % of reviews from CRCV corpus. The results are shown in Table 2. Following the previous experimental setting, 10 seed words were given for learning MANUFACTURE categories. We observed that the corpus′s size could have an influence on the results in Table 2. However, GMR-Bootstrapping achieved an exciting result with only 10 % of data. We also carried out another experiment to evaluate the impact of corpus′s quality. The last two rows in Table. 2 shows the results. 10 %*′s result was obtained from the selected 10 % CRCV reviews combined with 450 000 articles in other domains. The 10 %*′s result is much worse than the 10 %′s. The 20 %*′s result is similar to 10 %*′s. The results show that the percentage of the in-domain data is an important aspect for our methods. To our knowledge, it is not a difficult job to obtain the in-domain data from web.

Finally, we used the open domain Chinese corpus to evaluate the robustness of GMR-bootstrapping and Basilisk under the low quality corpus. The results are shown in Table 3. One can observe that GMR-bootstrapping′s performance is better than R-Basilisk′s in both categories. However, the performance is much worse than the results obtained under in-domain data. It is also proved that the quality of the corpus is an important aspect for both GMR-bootstrapping and Basilisk methods.

Table 2 GMR-bootstrapping in different sizes of Chinese CRCV corpus

Category	Percentage	GMR
MANUFACTURE	10 %	43
	20 %	56
	40 %	72
	80 %	96
	100 %	107
MANUFACTURE	10 %*	6
	20 %*	9

Table 3 GMR-bootstrapping vs. Basilisk in Chinese Open Domain Corpus

Category	Total words	GMR	R-Basilisk
IT Co. Name	100	17	9
	200	22	12
Fast Food Name	100	7	1
	200	12	1

3 Related work

Several weakly supervised classifier algorithms have been proposed to learn semantic lexicons with a small set of labeled data and a large number of unlabeled data, such as Co-training and Bootstrapping. The Co-training[1] alternately learns using two orthogonal views of data in order to use unlabeled data. This enables bootstrapping from a small set of labeled training data via a large set of unlabeled data. The KnowItAll[15] used a set of domain-independent extraction patterns to generate candidate facts. The candidate facts were evaluated by point-wise mutual information statistics. Snowball[6] used standard bootstrapping structure and introduced novel techniques for evaluating the quality of the patterns and tuples generated at each step of the extraction process. Hassan[16] presented an unsupervised method, which depends on redundancy in large data sets and graph-based mutual reinforcement to acquire extraction patterns.

The algorithm most closely related to our method is Basilisk[3], which is also a bootstrapping algorithm. While meta-bootstrapping trusts individual extraction patterns to make unilateral decisions. Basilisk gathers collective evidence from a large set of extraction patterns. We also used the same idea and structure, while there are some differences between GMR-bootstrapping and Basilisk. First, our method incorporates GMR to weight candidate words and extraction patterns. Second, we enhance the GMR-bootstrapping with pattern′s uncertainty to learn multiple categories simultaneously.

4 Conclusion and future work

In this paper, we present an approach to learn semantic lexicons using a new bootstrapping method, which is based on GMR. By changing the candidate words and patterns scoring functions, we incorporate GMR to weight the correctness of candidate words and extraction patterns. We

also enhance the GMR-bootstrapping to learn multiple categories simultaneously by adding patterns′ uncertainty into scoring functions. Another contribution of this work is that we present the impact of seed words′ quality on Basilisk and our method. Experimental results show that GMR-bootstrapping′s results are better than the previous best algorithm′s results in both of MUC4 and Chinese corpus.

References

1 Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory. Madison, USA: ACM, 1998. 92−100

2 Riloff E, Jones R. Learning dictionaries for information extraction by multi-level bootstrapping. In: Proceedings of the 16th National Conference on Artificial Intelligence. Orlando, USA: AAAI, 1999. 474−479

3 Thelen M , Riloff E. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: ACL, 2002. 214−221

4 Dong Z D, Dong Q. Hownet knowledge database [Online], available: http://www.keenage.com/, July 12, 1999

5 Miller G. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 1990, **3**(4): 235−244

6 Agichtein E, Gravano L. Snowball: extracting relations from large plain-text collections. In: Proceedings of the 5th ACM International Conference on Digital Libraries. San Antonio, USA: ACM, 2000. 85−94

7 Hirschman L, Light M, Breck E, Burger J D. Deep read: a reading comprehension system. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. Maryland, USA: ACL, 1999. 325−332

8 Moldovan D, Harabagiu S, Pasca M, Mihalcea R, Goodrum R, Girju R. Lasso: a tool for surfing the answer net. In: Proceedings of the 8th Text Retrieval Conference (TREC-8). Dallas, USA: TREC, 1999. 175−184

9 Riloff E, Schmelzenbach M. An empirical approach to conceptual case frame acquisition. In: Proceedings of the 6th Workshop on Very Large Corpora. Montreal, Canada: 1998

10 Florian R, Hassan H, Ittycheriah A, Jing H, Kambhatla N, Luo X. A statistical model for multilingual entity detection and tracking. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL. New York, USA: ACL, 2004. 1−8

11 Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extraction relations. In: Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. Barcelona, Spain: ACL, 2004. 178−181

12 Roark B, Charniak E. Noun-phrase cooccurence statistics for semiautomatic semantic lexicon construction. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational. Montreal, Canada: ACL, 1998. 1110−1116

13 Skounakis M, Craven M, Ray S. Hierarchical hidden Markov models for information extraction. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence. Acapulco, Mexico: Morgan Kaufmann, 2003. 1−7

14 Collins M, Singer Y. Unsupervised models for named entity classification. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Maryland, USA: ACL, 1999. 100−110

15 Etzioni O, Cafarella M, Downey D, Popescu A, Shaked T, Soderland S. Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 2005, **165**(1): 91−134

16 Hassan H, Hassan A, Emam O. Unsupervised information extraction approach using graph mutual reinforcement. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Austraili: ACL, 2006. 501−508

17 Riloff E, Wiebe J, Wilson T. Learning subjective nouns using extraction pattern bootstrapping. In: Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003. Edmonton, Canada: ACL, 2003. 25−32

18 Widdows D, Dorow B. A graph model for unsupervised lexical acquisition. In: Proceedings of the 19th International Conference on Computational Linguistics. Taipei, Taiwan: ACL, 2002. 1−7

19 Chinchor N. MUC-4 evaluation metrics. In: Proceedings of the 4th Conference on Message Understanding. Mclean, Virginia: ACL, 1992. 22−29

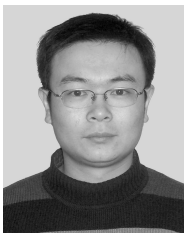
20 Zhang Q, Zhou Y Q, Huang X J, Wu L D. Graph mutual reinforcement based bootstrapping. In: Proceedings of Asia Information Retrieval Symposium. Harbin, China: Springer, 2008. 203−212

21 Riloff E , Shepherd J. A corpus-based approach for building semantic lexicons. In: Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing. Somerset, USA: ACL, 1997. 117−124

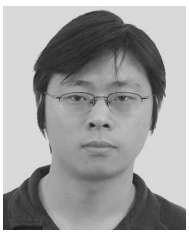
22 Riloff E. Automatically generating extraction patterns from untagged text. In: Proceedings of the 13th National Conference on Artificial Intelligence. Portland, Oregon: AAAI, 1996. 1044−1049

23 Kleinberg J M. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 1999, **46**(5): 604−632

24 Cover T M, Thomas J A. *Elements of Information Theory*. New York: John Wiley and Sons, 1991



ZHANG Qi Ph. D. candidate at the Department of Computer Science Engineering at Fudan University. His research interest cover sentimental classification and information extraction.
E-mail: qi_zhang@fudan.edu.cn



QIU Xi-Peng Assistant professor at Fudan University. He received his Ph. D. degree in computer science engineering from Fudan University in 2005. His research interest covers is machine leaning and image processing.
E-mail: xpqiu@fudan.edu.cn

HUANG Xuan-Jing Professor at Fudan University. She received her Ph. D. degree in computer science engineering from Fudan University in 1998. Her main research interest is natural language processing. Corresponding author of this paper.
E-mail: xjhuang@fudan.edu.cn

WU Li-De Professor at Fudan University. He graduated from Mathematic Department at Fudan University in 1958. His research interest covers natural language processing and image processing.
E-mail: ldwu@fudan.edu.cn