A Semi-Supervised Learning Method for Names of Traditional Chinese Prescriptions and Drugs Recognition

Dongfeng Cai, Changlin Ding, Junjun Zuo, Yu Bai
Research Center for Knowledge Engineering
Shenyang Aerospace University
Shenyang, China
e-mail: baiyu@sau.edu.cn

Abstract—**Knowledge discovery of Ancient Medical Literatures (AMLs) is a research focus due to wide applications of computer technology in Traditional Chinese Medicine (TCM). The foundation of the knowledge discovery research is to get semantic labels within the AMLs and to restructure the text. Due to the diversity of AMLs, low coverage rate of current semantic lexicons and the ambiguities of the lexicon words, low recall rate and low accuracy are resulted by using only lexicons to recognize TCM terms. This paper presents a semi-supervised learning and Bootstrapping based approach, Barpidusk, which aims at using semantic lexicons and simple features to recognize Names of Traditional Chinese Prescriptions and Drugs (NTCPDs) in un-annotated AMLs. And human-computer interaction is added to the Bootstrapping, which increases recognition accuracy without much loss of efficiency. Experiments show that the F values of recognizing Names of Traditional Chinese Prescriptions (NTCPs) and Names of Traditional Chinese Drugs (NTCDs) reaches 44.9% and 51.3% respectively without interaction with humans. By gradually adding human-computer interactions 10 times during recognition process, these values are increased to 74.9% and 90.6% respectively.**

Keywords-ancient medical literature; recognition for names of traditional Chinese prescriptions and drugs; Bootstrapping; human-computer interaction

I. INTRODUCTION

Semantic Annotation (SA) of Ancient Medical Literatures (AMLs) is the premise of lots of researches, such as associated rule mining and knowledge discovery. SA is based on the hypothesis that the named entities mentioned in documents constitute important part of their semantics [1].

Identifying references of entities in unstructured text is called Named Entity Recognition and Classification (NERC) [2]. According to the learning methods NERC used, NERC can be separated into 3 categories: supervised learning (SL), semi-supervised learning (SSL) and unsupervised learning (UL) based. Many Statistic Models (SMs) are used in SL [3, 4]. A typical approach in UL is clustering which cannot guarantee the recognition accuracy [2].

Large scale accurate initial training corpus (usually annotated manually) which is needed by SMs is hard to get. Furthermore, the training process of SMs is difficult to control and to understand; and users' intelligence is hard to be absorbed by SMs during the training process.

Bootstrapping is a framework for improving a learner using unlabeled data [5]. A corpus based Bootstrapping algorithm is designed in [6] to construct semantic lexicon. In [7], Mutual Bootstrapping and Meta Bootstrapping are proposed. And lots of researches are inspired from Mutual Bootstrapping [8, 9, 10]. All the methods mentioned above use large scale un-annotated corpus as input, and only recognize semantic categories within general domains.

Lexical and syntactic knowledge are used, so far, in Bootstrapping methods. But for Chinese processing, ambiguities and mistakes would occur from the phase of segmentation, and would be amplified by subsequent phases. For AMLs, it is more difficult to guarantee the accuracy of corpus processing then modern literature. Due to lack of lexical and syntactic knowledge in AMLs, current Bootstrapping methods are not fit for ancient literature.

In this paper, a method named Barpidusk, which uses simple context features instead of lexical and syntactic knowledge, is proposed. In the method, iteration process is simplified; and ambiguities or mistakes brought by pre-processing of corpus are avoided. Furthermore, human-computer interaction is added to Bootstrapping for generating a new framework, within which the semantic drift [10] is resolved and system efficiency is improved.

II. REVIEW OF BOOTSTRAPPING ALGORITHM

A. Initialization

Several seed words (usually 5) are selected for each semantic category in accordance with the principle: the seed words must be frequent in the domain and unambiguous [6].

B. Iteration

1) Seed words are used for generating more patterns

Both Mutual and Basilisk assume that the quality of a pattern is measured by two factors which are the pattern's reliability and the pattern's efficiency for extracting words. The factors can be expressed in (1)[7, 8]:

$$score(pattern_i) = F_i / N_i \times \log_2(F_i) \quad (1)$$

Where F_i is the number of seeds extracted by $pattern_i$; and N_i is the total number of words extracted by $pattern_i$.

2) Patterns are used to extract more candidate words

Basilisk Bootstrapping assumes that if a pattern had extracted many category members, it tended to extract more, so the words extracted by these patterns can earn a higher score. The assumption can be formalized as (2)[8]:

$$score(word_i) = {}^{1}\!/_{P_i} \sum_{j=1}^{P_i} \log_2 (F_j + 1) \quad (2)$$

Where P_i is the number of patterns that extract $word_i$; and F_j is the number of distinct category members extracted by $pattern_j$.

C. Termination conditions

- When loop count equals a given value of loop times, Bootstrapping ends [6, 7, 8];
- When the maximum score of the patterns or words are less than a threshold, Bootstrapping ends [7];

III. Recognition for Names of Traditional Chinese Prescriptions and Drugs

An approach called Barpidusk (Bootstrapping Approach to Recognize Names of Traditional Chinese PrescrIptions and DrUgs using Semantic Knowledge) is proposed.

A. Barpidusk Bootstrapping

Fig. 1 shows the steps of Barpidusk. A value of loop times is set as termination condition and two kinds of seed words are contained in the input of this approach.

- *Correct seed words*: In the field of Traditional Chinese Medicine (TCM), many lexicons of TCPs and TCDs have already been generated. Words in both lexicon and corpus are selected as initial correct seed words instead of manually selected seed words.
- *Incorrect seed words*: The function of these words is just as that of words in the stop word list. With these words, loop efficiency is increased and wisdom of users can be added at the beginning of the loop.

B. Generating Patterns

Context instead of phrase information is used for pattern generation. In order to obtain more general and multilayer patterns, a 3 size window is used for capturing context. And "dimension" is used to represent the length of each pattern.

Some pattern generation principles are shown as follows:

- In order to prevent over spreading of patterns, punctuations which signify end of sentence, such as period, question mark et al can not be surpassed.
- In order to assure context integrality within every pattern, the number of words before and after seed word in each pattern must not be less than 1.
- Regular expressions are used in generated patterns. More specifically, regular expression "(.+?)" is to replace the seed word, where "()" is the group wanted to be extracted; ".+" represents one or more random characters; "?" means un-greedy extraction.

Table 1 demonstrates an actual example, the sentence "……薛作风客淫气，治以**地黄丸**而愈。……" is selected from an AML and the bold word in sentence is a NTCP.

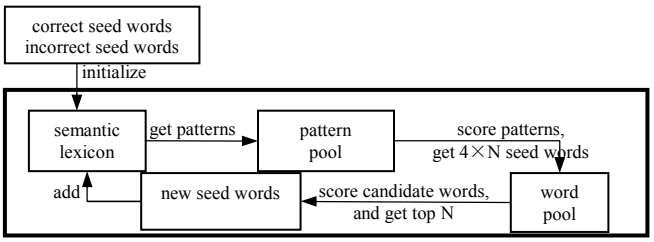

Figure 1. Barpidusk

TABLE I. Instance of Building Patterns

Dimension	**Patterns**(*Separated by* ||)
2	以**(.+?)**而
3	以**(.+?)**而愈 ‖ 治以**(.+?)**而
4	治以**(.+?)**而愈 ‖ 以**(.+?)**而愈。 ‖，治以**(.+?)**而
5	，治以**(.+?)**而愈 ‖治以**(.+?)**而愈。
6	，治以**(.+?)**而愈。

C. Obtaining Candidate Words in Barpidusk

1) Extracting candidate words in Barpidusk

Without lexical and syntactic knowledge, regular expression may result in lots of problems. For example, if pattern "以(.+?)而" is used upon sentence "以致痰涎上涌，火载血而上行", "致痰涎上涌，火载血", which cannot be a NTCPD, will be extracted. In order to avoid this kind of obvious errors, following filter principles are established.

a) Punctuation filtering: Generally, no punctuations, except a few ones as in "参、芪大补之剂" et al, are in NTCDs and NTCPs. So all the extrated words which contain punctuations are filtered directly;

b) Length limitation: In order to filter the words that have too short or too long lengths, a range of average length of current seed words ±1 is defined and all the candidate word lengths must be within this range.

2) Improving candidate word scoring equation

According to (2), for example, pattern P_1 extracted 100 words and half of them are correct, pattern P_2 extracted 50 words and 40 of them are correct, if candidate words S_1 and S_2 fit P_1 and P_2 respectively, according to (2) of the assumption, Score(S_1) = $\log_2 51$ = 5.67, Score(S_2) = $\log_2 40$ = 5.32. However, P_2 is obviously more reliable than P_1. So, the assumption of (2) is improper and (2) is modified to (3):

$$score(word_i) = {}^{1}\!/_{P_i} \sum_{j=1}^{P_i} \log_2 \left({}^{F_j}\!/_{N_j} \times 10 + 1 \right) \quad (3)$$

Where, P_i and F_j have the same meaning with the ones in (2) and N_j is the number of words P_i extracted. According to (3), Score(S_1)=$\log_2 6$=2.58 and Score(S_2)=$\log_2 9$=3.17. The rationality of (3) will be proved by following experiments.

IV. The Human-Computer Interaction Based Barpidusk

In this paper, human-computer interaction is applied three times in the process of Barpidusk as shown in Fig. 2.

A. Classifing candidate words

Manual judgment 1 in Fig. 2 denotes this process. Users' wisdom undoubtedly can improve the recognition accuracy. It is proposed in [7] the performance of Bootstrapping will be better if the results of loops can be modified manually.

Except the extracted candidate words which are themselves NTCPDs, the remaining ones which either contain a NTCP or be contained by a NTCP are also valuable for users. So, three sets are defined, set contains matched actual NTCPDs, set contains modifiable candidate words and set contains incorrect candidate words. After given times of loops (inner-loop in Fig. 2), every candidate word would fall into one of the above sets according to judgment of user, and put into the equivalent seed word set for next group of loops.

B. Finding out NTCPDs through modifiable words

This process is the manual judgment 2 shown in Fig. 2. Users can select the corresponding NTCPDs for modifiable words among the words that are recommended by system; these selected NTCPDs are then put into correct seed word set for next group of loops.

C. Filtering correct seed words

This process is denoted as manual judgment 3 in Fig. 2. If the initial correct seed words are not judged at the beginning of the every group of loops, the seed words will be ambiguous, and semantic drift would occur immediately.

V. EXPERIMENT RESULT

"Classified Medical Records of Distinguished Physicians" (名医类案) and "Classified Medical Records of Distinguished Physicians Continued" (续名医类案) are used as corpus of experiments which contain 519 medical records, 247 NTCPs, 352 NTCDs, 3878 sentences and 93858 words (without punctuations).

A. Experiments

Group 1 (G1): Aim to prove the validity of Barpidusk and find out the best parameter combination.

Group 2 (G2): Simulant human-computer interaction is added into Barpidusk with the best parameter combination obtained from G1.

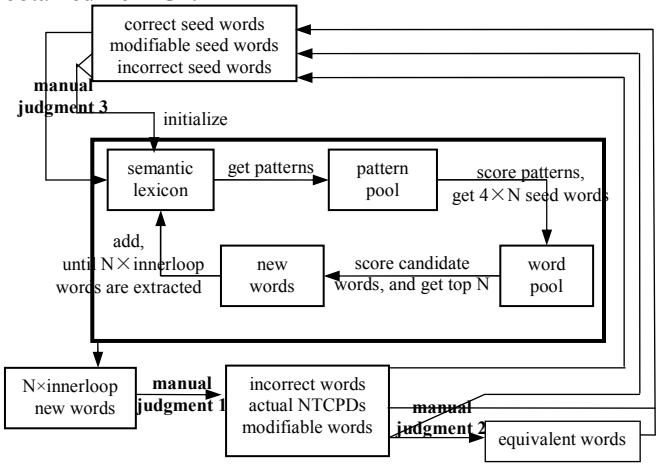

Figure 2. Barpidusk with Human-Computer Interaction

Some NTCDs appear in the similar context with NTCPs, and many NTCDs are contained within NTCPs. To avoid category cross, first, NTCPs are recognized; second, all correctly recognized words in corpus are deleted; and finally, new corpus is used for extracting NTCDs.

B. G1

In G1, the selected seed words for initialization are comprised by two parts: 51 NTCPs and 207 NTCDs. Fig. 1 shows the process of G1 with N set to be 5.

1) Parameters

Except the equations used to score candidate words, following alternative parameters need to be examined.

a) Incorrect seed words: To avoid extracting semantic words of another category, the 207 NTCDs are treated as incorrect seed words for NTCP recognition and correctly recognized NTCPs are used as incorrect seed words for subsequent recognition of NTCDs.

b) Reprocessing of NTCPs

- *Integration of NTCP:* Some extracted candidate words are parts of equivalent NTCPs, these words may become NTCPs by adding one of the following suffixes "汤", "散", "丸", "丹", "膏", "饮".
- *Filtration of NTCPs:* Some phrases which contain fixed strings or end with fixed strings appear in the similar context with NTCPs, but they are not NTCPs. To directly filter these candidate words, two string sets are established which are shown in Table 2.
- *NTCD + suffix = NTCP:* A NTCP can be made up of a NTCD and a suffix..

2) Recognition of NTCPDs

Table 3 shows parameters of 4 experiments for NTCP recognition and Table 4 shows the ones of 3 experiments for NTCD recognition. The results are shown in Fig. 3 to Fig. 4.

3) Analysis of G1

Obviously, the performances of the fourth experiment for recognizing NTCPs and the third for NTCDs are the best.

No obvious difference is observed through comparison between No.1 and No.2 of NTCPD recognition. This is for following reasons, first, the incorrect seed words can only filter a part of incorrect words; second, even the incorrect candidate words are filtered, the accuracy of added words to substitute words that are filtered cannot be guaranteed.

Seen from the comparison between No.2 and No.3 of NTCP recognition, the reprocessing is helpful obviously. However, same problem mentioned above occurs. That is, the accuracy of substituting words cannot be guaranteed.

Last two experiments for NTCPD recognition are compared. By applying (3), recognition effect is increased.

TABLE II. TWO SETS OF FIXED STRINGS FOR NTCP RECOGNITION

String Set	Strings	Instances
Fixed strings that are contained by incorrect phrases	number + units in TCM("分", "两" etc.) special strings	"两钱" "之药"
Fixed strings that are ends of incorrect phrases	"药"	----

TABLE III. PARAMETERS FOR NTCP RECOGNITION IN G1

NO.	Word Scoring Equation	Incorrect Seed Words	Reprocessing of NTCPs
1	(2)	no	no
2	(2)	yes	no
3	(2)	yes	yes
4	(3)	yes	yes

TABLE IV. PARAMETERS FOR NTCD RECOGNITION IN G1

NO.	Word Scoring Equation	Incorrect Seed Words
1	(2)	no
2	(2)	yes
3	(3)	yes

C. G2

Labeled NTCPDs in the corpus are to be compared with the candidate words obtained after every group of loops. Through the comparison, the candidate words are classified into three sets defined before. Without human intervention, manual judgment 2 and 3 cannot be achieved.

The process showed in Fig. 2 can describe the steps of G2 with manual judgment 2 and 3 omitted and with manual judgment 1 replaced by the comparison introduced before.

As long as 50 new candidate words are extracted, next group of loops would be activated.

Because the candidate words are modified after every group of loops, accuracy, recall rate and F value will have undulated improvement with the increasing times of loops. Here in our experiments, when 500 candidate words (after 10 groups of loops) are extracted, the best performance of G2 appears as Table 5 shows.

1) Analysis of G2

After 10 groups of loops, only 58 NTCPs and 28 NTCDs are not extracted. All the NTCPDs that are not recognized are low frequency words($\leqslant$3), and 91.32% of them appear only once in corpus. Except the small scale of corpus, aliases or abbreviations of NTCPDs and irregularity of NTCPDs also cause low frequencies of NTCPDs.

VI. CONCLUSION

The traditional Bootstrapping algorithms are adjusted and improved in this paper, and a new method, Barpidusk, is proposed for recognizing NTCPDs in AMLs. Experiments prove that Barpidusk, which generates patterns from simple context instead of lexical and syntactic knowledge, is feasible. Furthermore, the addition of human-computer interaction into Barpidusk Bootstrapping also results considerable performance.

Future work will be focus on testing this method by other more corpuses and extending this method into recognition of other semantic categories within TCM or other domains.

ACKNOWLEDGMENT

Our work was supported by 973 Program of China, 2010CB530401.

TABLE V. BEST PERFORMANCE FOR NTCPD RECOGNITION OF G2

NTCP	Actual NTCPD	Modifiable Words Included
Accuracy	0.682	0.734
Recall Rate	0.425	0.765
F Value	0.524	0.749
NTCD	**Actual NTCPD**	**Modifiable Words Included**
Accuracy	0.85	0.893
Recall Rate	0.724	0.920
F Value	0.782	0.907

REFERENCES

[1] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, and Damyan, "KIM-semantic annotation platform," Proc. ISWC, Spinger Press, Florida, USA, 2003, pp. 834-849.

[2] D. Nadeau and S. Sekine, "A survey of name entity recognition and classification," Lingvisticae Investigationes, vol. 30, 2003, pp. 1-20.

[3] M. Asahara and Y. Matsumoto, "Japanese named entity extraction with redundant morphological analysis," Proc. HLT-NAACL, May 2003, pp. 8-15.

[4] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," Proc. CoNLL, 2003.

[5] R. Jones, A. McCallum, K. Nigam, and E. Riloff, "Bootstrapping for text learning tasks," Proc. IJCAI, Sweden, 1999, pp. 52-63.

[6] E. Riloff and J. Shepherd, "A corpus-based bootstrapping algorithm for semi-automated semantic lexicon construction," Natural Language Engineering, vol. 5, 1999, pp. 147-156.

[7] E. Riloff and R. Jones, "Learning dictionaries for information extraction by multi-level bootstrapping," Proc. Sixteenth National Conference on Artificial Intelligence, 1999, pp. 474-479 .

[8] M. Thelen and E. Riloff, "A bootstrapping method for learning semantic lexicons using extraction pattern contexts," Proc. EMLNP, Philadelphia, USA, July 2002, pp. 214-221.

[9] M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain, "Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge," Proc. National Conference on Artificial Intelligence, AAAI Press, 2006, pp. 1400-1405.

[10] J. R. Curran, T. Murphy, and B. Scholz, "Minimizing semantic drift with mutual exclusion bootstrapping," Proc. PACLING, Melbourne, Australia, 2007, pp. 172-180.

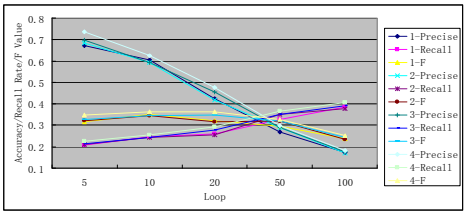

Figure 3. Comparison of NTCP Recognition of G1

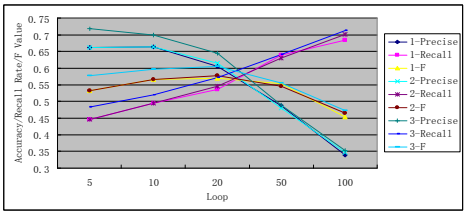

Figure 4. Comparison of NTCDs Recognition of G1