Collaborative Recognition and Recovery of the Chinese Intercept Abbreviation*

Jinshuo LIU[1], Yusen CHEN[1], Juan DENG[2+], Donghong JI[1], Jeff PAN[3]

[1] Computer School, Wuhan University, Wuhan 430072, China
[2] International School of Software, Wuhan University, Wuhan 430072, China
[3] University of Aberdeen, Aberdeen, AB24 3FX, UK
dengjuan@whu.edu.cn

Abstract. One of the important works of Information Content Security is evaluating the theme words of the text. Because of the variety of the Chinese expression, especially of the abbreviation, the supervision of the theme words becomes harder. The goal of this paper is to quickly and accurately discover the intercept abbreviations from the text crawled at the short time period. The paper firstly segments the target texts, and then utilizes the Supported Vector Machine (SVM) to recognize the abbreviations from the wrongly segmented texts as the candidates. Secondly, this paper presents the collaborative methods: Improve the Conditional Random Fields (CRF) to predict the corresponding word to each character of the abbreviation; To solve the problems of the 1:n relationship, collaboratively merge the ranking list from the predict steps with the matched results of the thesaurus of abbreviations. The experiments demonstrate that our method at the recognizing stage is 76.5% of the accuracy and 77.8% of the recall rate. At the recovery step, the accuracy is 62.1%, which is 20.8% higher than the method based on Hidden Markov Model (HMM).

Keywords: Collaborative Recovery; Improved CRF; Chinese Abbreviation

1 Introduction

The most important information content security supervision methods is to match the sensitive words and rules of the text, which needs to manually or semi- automated build up the corpus and the thesaurus of rules. There are 2 challenging tasks for supervision. ① To avoid the supervision and management, the theme words or kernel phrases are sometimes substituted by the abbreviation, which cannot be found through preprocessing of the data. ② The manually built up corpus although is accurate, but needs to be continuously updated. So it is another tough job to quickly automated update the corpus via recognizing the recovering the related phrases from the relatively texts crawled in a short time.

[*] + Corresponding author: E-mail: dengjuan@whu.edu.cn

This paper is supported by National Science Foundation of China (NOs.61672393, U1536204)

The abbreviation is one of the most important forms to substitute the original phrases. So the recognition and recovery of the abbreviation is very important to public supervision. Because of the complexity of the Chinese and difference with English, the research for Chinese abbreviation is hard.

Chinese abbreviation means intercepting, abridging, concluding, changing the order of the original length of word without changing the meaning [1]. There are three type of abbreviation, intercept abbreviation, abridged abbreviation and concluding abbreviation. The intercept abbreviation only keeps the kernel morphemes to represent the original phrase. For example, "北大" (BěiDà) means the "北京大学"(Beijing University),which is the most important ways to avoid the supervision.

This paper is to quickly and accurately discover the intercept abbreviations from the texts crawled in the short time period. The paper firstly recognizes the abbreviations via SVM from the wrongly segmented texts. Secondly, this paper presents the collaborative recovery methods of the abbreviation. Predict the corresponding word to each character of the abbreviation with the improved CRF. The abbreviation—originate lookup table can also be updated finally.

The structure of the paper is listed as followings: Part 1 introduces the challenges of the Chinese abbreviation for network security. Part 2 is the related research work. Part 3 is our main methods. Part 4 introduces the experiments including performance experiments and comparison experiments. Part 5 is the conclusion.

2 Related Works

The recognition and recovery of abbreviations cannot be separated from the study of informal words because the abbreviations account for about 20% of informal words [2].

In study of the informal word recognition, Wang et al. [2] employ a factorial conditional random field to model both tasks of informal word recognition and Chinese word segmentation jointly. Besides that, Wang et al. [3] also use a large-scale corpus to select the formal equivalents of informal words according to context semantic similarity, and then formalize the task of informal word recovery as a binary classification problem. This method is feasible but can't achieve a rather high accuracy. Li et al. [4] classify the words into three classes: IV (in-vocabulary) correct-OOV（out-of-vocabulary) and ill-OOV, and proposed a non-standard word detection method based on the maximum entropy classifier. Monroe et al. [5] extend an existing Modern Standard Arabic segmenter with a simple domain adaptation technique and new features in order to segment informal and dialectal Arabic text.

For the recovery of abbreviation, Chang et al. [7] consider the original word is in a hidden state. And they map this problem to a HMM. Although, this method has a good result in intercepted abbreviations, it doesn't utilize the contextual features well. Roack et al. [8] focus on abbreviation recovery for text-to-speech synthesis and they use an n-gram model and SVM model to classify the candidate expansion of the abbreviation.

In study of the abbreviation prediction, Yan Jiao et al. [9] use conditional random field to generate a number of candidates, then they re-score the candidates according to the results from web search engine. And the one with the highest score is selected as the abbreviation. Kailong Zhang et al. [10] introduce the minimum semantic unit to capture word level information based on the CRF and use an integer linear programming formulation to recode the abbreviation from the generated candidates. Besides that, they also propose a two-stage method [11] to find the corresponding abbreviation. First they use a large scale corpus to generate candidates and get a coarse-grained ranking through graph random walk. Then re-rank the candidates according to the feature. Chen H et al. [12] propose a novel abbreviation generation method using first-order logic and markov logic network frameworks. Yangyang shi et al. [13] use a RNN model with maximum entropy extension in abbreviation prediction and generation.

All approach mentioned above builds on an important part: generate an abbreviation lookup table. Although this method can build a corpus quickly and efficiently, it may have high false detection rate for the reason of abbreviations' ambiguity.

3 Collaborative Recognition and Recovery

The paper devises three stages for recognizing and recovering the Chinese Abbreviation: pre-process, recognition and recovery. The framework of our method is listed as Fig. 1.

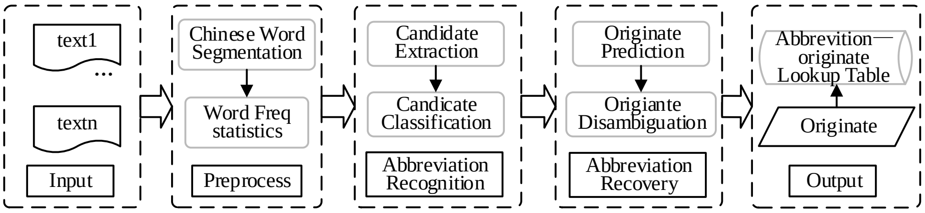

Fig. 1. The framework of abbreviation collaborative recognition and recovery

- **Preprocess.** The text need to be segmented into a word sequence. As the unlogged words, abbreviations cannot be correctly segmented, but be left as the text slices or wrongly cut into 1-gram sequences.

Recognition. This paper selects the 1-gram sequences or long text slice as the candidate abbreviations and utilizes the information and context of the abbreviation as the features. The abbreviation can be classified into 2 sets, 'yes' or 'no', using SVM. 'yes' is the true abbreviations.

Recovery. This paper converts the problem of recovering the abbreviation into the prediction of each character. The traditional abbreviation-originate lookup table is utilized to assist in the originate disambiguation. The final originates are gotten through the the cosine similarity computation of candidate originates.

3.1 Rule and context based abbreviation recognition

This paper utilizes two filtering stage for recognizing the abbreviations. The first stage extracts the candidate abbreviations from the unlogged text slices. The second stage uses the SVM classifier to classify the candidates into two classes according to the features of rules and context. (Shown as Fig. 2)

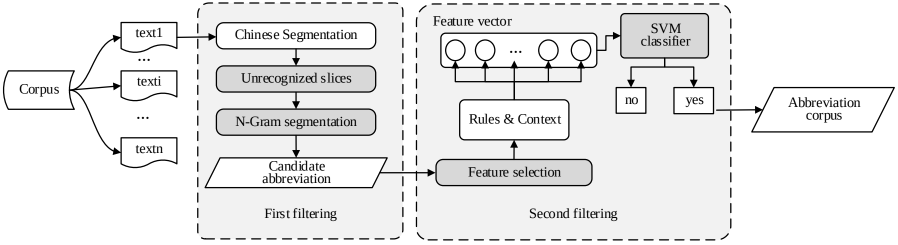

Fig. 2. Rule and context based recognition of abbreviation

Extraction of the unrecognized slices. Abbreviation as the unlogged words cannot be correctly segmented finely as the word sequences. After preprocessing the corpus, select the incorrectly segmented slices or consecutive unigram as the candidate abbreviation.

N-gram segmentation. Normally speaking, the abbreviation is 2 to 4 characters long. The ratio of length of abbreviation over 4 is very small. The single character abbreviation basically represents the capital in Chinese. So this paper utilizes the 2-4 characters long abbreviations as our candidates. Since the initial candidates normally are long, and need to be segmented further, we use the N-gram segmentation method to segment the candidates, and select 2 or 4 words as the final candidates. For example, “计生委” as the coarse candidate, and further segmented as “计生”, “生委” and “计生委”。

Feature selection. This paper selects following three types of features:

Information feature. Define the abbreviation $x_1^n = x_1 x_2 \dots x_n$ as the sequences with length *n*, $2 \le n \le 4$.

- *Length m*. According to the statics, the proportion of bi-gram type is 55%, tri-gram is also normal, and quad-gram is the least. So we set the length as one feature.
- *Number.* numbers and the positions of them are also import discriminate features in Chinese. Normally speaking, number is very important to record the related words in the abbreviation. For example, “二” (second) of “第二次世界大战” (the Second World War) is so important, and is kept as “二战”.
- *Prefix and suffix.* In the abbreviations, some Chinese characters are used as the start and end of the abbreviations. For examples, “局” is the end of the abbreviation such as “国安局” etc. We utilize the prefix and suffix with frequency as features and represented as ‘word bag model’.
- *The probability of the positions* $\mathrm{P_{pos}}(\mathrm{x_1^n})$.

$$P_{pos}(x_1^{\mathrm{n}}) = \prod_{x \in X} P(pos, x) \tag{1}$$

Where $P(pos, x)$ is the probability vector of character x at position *pos*. We limit the the value of *pos* to begin, middle and end. In Chinese, some characters "局、会、部" often appear at the end of the abbreviations. Some characters such as "中、北、东" often appear at the start or the middle of the abbreviations. So the higher one unpredicted abbreviation is, the more possible a real abbreviation is.

- Morph-abbreviation. Suppose the abbreviation $x_1^3 = x_1x_2x_3$is unified by x_1, x_2, x_3, which are from the logged word set $\{x_1x_2, x_1x_3\}$ or$\{x_1x_3, x_2x_3\}$ [9]. The whole type of the word is unified with multiple same prefixes or suffixes. For example, "中医、西医" merge as "中西医".

Context feature. The context of the target object normally includes enough information. Define the slice text $W_{i-2}W_{i-1}x_1^nW_{i+1}W_{i+2}$.

- *The frequency of the abbreviation tf.* If an abbreviation is a real abbreviation, then the occurrence of them is definitely higher than other candidates.
- *Bigram information entropy.* The entropy of the bigram information entropy is represented by:

$$BH2(x_1^n) = -\sum_{w \in W} \frac{C(w,x_1^n)}{n_c} \log \frac{C(w,x_1^n)}{n_c} \tag{2}$$

Where n_c is the occurrence of the abbreviation x_1^n, W is the left or right word set of x_1^n， $C(w, x_1^n)$ is the co-occurrence of word *w* and x_1^n. The left and right entropy of the words evaluate the frequency of the left and right words. [14] The entropy can be bigger and bigger with the frequency of the combination of the words get higher and higher.

- *Tri-gram information entropy.* It is the information entropy of combination words, $W_{i-2}W_{i-1}x_1^n$, $W_{i-1}x_1^nW_{i+1}$ and $x_1^nW_{i+1}W_{i+2}$

Global features. The higher global information a candidate abbreviation holds, the more possible a real abbreviation is.

- *Frequency of the abbreviation in the corpus tfc.*
- *The number of the text containing the abbreviation idf.*

Classifier. We choose the Support Vector Machine (SVM). SVM is based on the least risk and more suitable for small sample set. Since the abbreviation is sparse with small training set.

3.2 Multi-label Sparse Conditional Random Field based Recovery

After the abbreviation has been recognized, we need to recover it to the corresponding originate. Using the feature that the abbreviations and originates have the 1:1 mapping relationship, we convert the problem of recovering the abbreviation into the prediction of each character and achieve the word-sequence prediction based on multi-label sparse conditional random field (SCRF). If the probability of all prediction results are low, we obtain the matched originate from the abbreviation-originate lookup table and put it together with the predicted originates as originate candidates. All candidates is ranked through the cosine similarities to obtain the final originate that is best for the current context to achieve the abbreviation disambiguation. Finally, record the new originates to extend the lookup table. The concrete process is shown in Fig. 3.

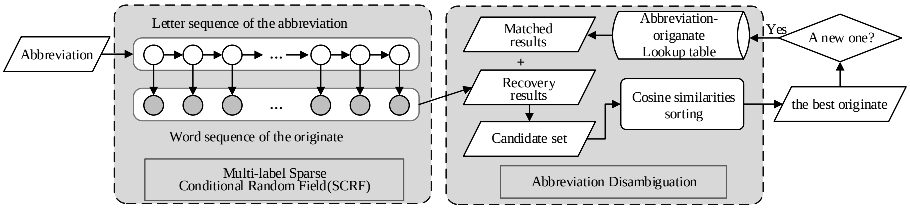
Abbreviation
Letter sequence of the abbreviation
...
Word sequence of the originate
Multi-label Sparse Conditional Random Field(SCRF)
Matched results
+
Recovery results
Candidate set
Abbreviation-origanate Lookup table
Cosine similarities sorting
Abbreviation Disambiguation
Yes
A new one?
the best originate

Fig. 3. The abbreviation recovery based on multi-label sparse conditional random field

Multi-label sparse conditional random field. This paper adopts the linear chain conditional random field model to achieve the abbreviation recovery. Take the abbreviation sequence x_1^n as the input sequence, and $x_i \in X$ is the Chinese character set. Take the originate sequence y_1^n as the output sequence, and $y_i \in Y$ is the word set. The conditional probability distribution $P(y_1^n|x_1^n)$ of the sequence y_1^n constitutes the conditional random field which satisfies Markov property

$$P(y_\mathrm{i}|x_1^n, y_1, y_2 \dots y_n) = P(y_\mathrm{i}|x_1^n, y_{i-1}, y_{i+1}) \tag{3}$$

Then we call $P(y_1^n|x_1^n)$ the linear chain conditional random field. Under the given input sequence x_1^n, the conditional probability distribution of the output sequence y_1^n is as follows:

$$P(y_1^n|x_1^n) = \frac{1}{Z(x_1^n)} exp(\sum_{t=1}^{n} \sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t,)) \tag{4}$$

$$Z(x_1^n) = \sum_y exp(\sum_{t=1}^{n} \sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t,)) \tag{5}$$

In above formulas, f_k is the feature function, θ_k is the corresponding weight, and $Z(x_1^n)$ is the normalization factor. From the Formula (4), we can see that the number of the feature function is related to the state set X and the label set Y, and the parameter K satisfies the equation:

$$K = |Y|^2 \times |X|_{train} \tag{6}$$

$|X|_{train}$ denotes the number of all Chinese characters of the abbreviations in the training set, and $|Y|$ is the number of all words of the corresponding originates. It is obvious that X and Y contain a large number of elements, so the scale of the feature function set $\{f_k\}$ will be extremely large, which makes the learning difficulty of the CRF model grow exponentially.

This paper adopts a series of methods to simplify the model learning process, the main goal of which is to reduce the scale of the feature weight set $\{\theta_k\}$ and then to obtain the improved multi-label SCRF model.

(1) The first method of simplifying the model is aimed at reducing the number of the elements in the feature function weight set $\{\theta_k\}$. Given the set $\phi(x,y) \in M$, and $\phi(x,y): x \rightarrow y$ denotes the mapping relationship between the Chinese character *x* and the word *y*. In the pair <abbreviation, originate>, if the Chinese character x of the abbreviation has a corresponding relationship with the word y of the originate, then the mapping relationship $\phi(x,y)$ holds. It is easy to know that after a Chinese character in the abbreviation set is recovered, the corresponding word set Y_x only contains a few words and there is no relations between the other words in the set Y. Therefore, given $x \in X$ and $y \in Y$, if there is no mapping relationship between x and y in the training set, then the features of *x* and *y* will be ignored which is achieved by setting the corresponding feature weights as $-\infty$, as shown in Formula (7).

$$\theta_k(y_i, x_j, i) = -\infty, \theta_k(y_i, x_j, i) \in \{\theta_k\}, s.t.\ \phi(x,y) \notin M \tag{7}$$

(2) The second method of simplifying the model is aimed at reducing the computational complexity of the model learning. This paper adopts the regularization term ℓ_1 to enhance the sparse degree of the model. This method tends to create less but more useful features in the model and lets most of the feature weights be zero, which will reduce the memory usage and optimize the forward-backward calculation in the model, and it greatly simplifies the model learning process to some degree.

Therefore, this paper takes the improved SCRF model as the abbreviation recovery method at last, and adopts the block-coordinate descent method [17] to achieve the model learning process.

Feature Selection. This paper selects features according to the related information of each Chinese character of abbreviations when recovering the abbreviations. If the abbreviation sequence $x_1^n = x_1 x_2 \ldots x_n$ is given, and x_i denotes the *i*th Chinese character of the abbreviation, then the selected features are as follows:

- *The Chinese character and Chinese phonetic alphabet of* $\mathrm{x_i}$ *and* $\mathrm{x_{i-1}}$. The pronunciation of the Chinese character is related to its corresponding originate, which means that the Chinese character in the corresponding word has the same pronunciation as the original Chinese character in the abbreviation. For example, in the abbreviation “中行” (BOC) and in its corresponding originate “中国银行”(BANK OF CHINA), the two “行”s have the same pronunciation “háng” in the Chinese phonetic alphabet.
- *The word with the highest frequency in all short words that contain the Chinese character* $\mathrm{x_i}$ *in the text, its length and phonetic alphabet.* An abbreviation can be recovered to several different originates under different contexts. For example, the abbreviation “南大”(NU) can be recovered to “南京大学”(Nanjing University) and “南昌大学”(Nanchang University). It’s found in researches that as for this kind of abbreviations, the corresponding word of the polysemantic Chinese charac-

ter tends to appear in the context several times to help readers disambiguate, and its word frequency is higher than other words that contain the same Chinese character.

- *Whether* $\mathrm{x_i}$ *and* $\mathrm{x_{i-1}}$ *is a number or not.*
- *Morph-abbreviation*.

Abbreviation Disambiguation. There is a one-to-many relationship between the abbreviations and corresponding originates. Sometimes all the prediction results have a lower probability than 0.6, which means the best originate may be not in the prediction results. In this case, this paper chooses the top five originates with the highest prediction probabilities as the recovery candidate set. At the same time, according to the abbreviation-originate lookup table, we obtain the matching candidate set by matching the abbreviation in the lookup table. All the candidate originates will be evaluated according to the cosine similarity of context. We choose the sentence of the target abbreviation as context of candidate originates. The originate with highest score is the best result. Finally, add the new corresponding relationship to the dictionary, which automatically extends the dictionary. The calculation method of the cosine similarity is as follows:

$$\cos(context1, context2) = \frac{\sum_{i=1}^{n} c_{1i} \times c_{2i}}{\sqrt{\sum_{i=1}^{n} c_{1i}^2} \times \sqrt{\sum_{i=1}^{n} c_{2i}^2}} \tag{8}$$

4 Experiment and Analysis

4.1 Dataset

Abbreviations dataset. The abbreviated dataset used in this paper is mainly obtained in three ways, the Institute of Computational Linguistics, Peking University[1], Baidu Encyclopedia and the microblogging corpus

we get an abbreviation-originate lookup table which includes 14372 pairs of abbreviations-originate. Table 1 represents the proportion of each type of abbreviations and we choose the intercepting abbreviation as the dataset for this experiment.

Table 1. Comparison of the abbreviations of each category

	intercepting	abridging	concluding
PKU	5846	1679	634
Baidu	3527	1034	355
Microblog	985	234	78

Corpus dataset. In this paper, we select about 20000 documents containing abbreviations from the text classification corpus released by Sogou Lab[2] as the corpus dataset of our experiment.

[1] http://www.icl.pku.edu.cn/icl_groups/corpus/

[2] http://www.sogou.com/labs/resource/

4.2 Experiment Setup

Experiment 1: Performance experiment.

We divide the procedure of Chinese abbreviations into two problems, the recognition and the recovery of abbreviations, and perform two independent performance experiments.

Abbreviation recognition. This paper uses the NLPIR[3] tokenizer of the Chinese Academy of Science to implement the Chinese word segmentation. We selected the decision tree (DT) model, the logical regression (LR) model and the support vector machine (SVM) model as the binary classifier and use 10-fold cross validation to train the abbreviated recognition model.

Abbreviation recovery. Experiment preprocesses about 10,000 pairs of abbreviations – originates, which form a 1-to-1 word mapping for each character in the abbreviation. In the stage of feature extraction, because the eigenvalue of CRF can only be discrete, we use the equal frequency method to discretize the continuous eigenvalue. Finally, 10-fold cross validation is used in the training of the abbreviated recovery model.

Experiment 2: Comparison experiment.

To validate the universal property of our model, we compare our abbreviation recovery method based on the sparse condition random field with the abbreviation recovery method based on the hidden Markov model [7]. We implement two methods and conduct the comparison experiment over the same dataset.

4.3 Results

4.3.1 Performance experiment

The performance of the abbreviation recognition algorithm and the abbreviation recovery algorithm are verified respectively.

Abbreviations recognition method performance experiments

The DT, LR and SVM are compared shown in Table 2. To validate different kernel functions in SVM model, the Linear, POLY and RBF kernel are also compared. The experimental results of the different kernel functions are shown in Table 3.

It can be seen from Table 2 that the SVM model has a better effect than the LR model and the DT model. SVM model has higher improvement in precision than the other two models, but the LR model is similar to SVM model in recall rate and F value. In addition, it can be seen from Table 3 that although the SVM model with RBF kernel has higher precision, the SVM model with linear kernel is similar in precision to the other two kernel functions. Considering the data factors, the three kernel functions are not much different in terms of performance, which may be related to the higher feature dimension of the experiment.

[3] NLPIR: http://ictclas.nlpir.org/

Table 2. The experimental results of different classifiers in abbreviation identification methods

Classifier	Precision	Recall	F-value
DT	0.738	0.748	0.743
LR	0.734	0.769	0.751
SVM	**0.765**	**0.778**	**0.771**

Table 3. The experimental results of different kernel functions in SVM model

Kernel	Precision	Recall	F-value
Linear	0.761	0.771	0.766
Polynomial	0.758	0.766	0.762
RBF	**0.765**	**0.778**	**0.771**

Abbreviations recognition method performance experiments.

We divide the selected features into basic and extended features. The basic feature is all the features except the phonetic symbols in Section 3.4. The extended feature is the phonetic alphabet of each Chinese character in the abbreviation. The experimental results are shown in Table 4. Top-n indicates whether the correct results are included in the first n most likely results of the selected model.

Table 4. The results of abbreviations recovery experiments based on improved CRF

Features	Top-N	Precision
basic features	1	0.617
	2	0.688
	all	0.934
basic features + extended features	1	**0.621**
	2	**0.691**
	all	**0.934**

As can be seen from Table 4, the precision of the CRF model using basic features has been relatively high. In the basic features of the combination of expansion features for the precision has no significant improvement. This paper consider the expansion features because the Chinese characters have multiple pronunciations in different words, which will affect the meanings of the words. Therefore, we believe that the phonetic feature can improve the prediction ability of CRF model. Experiments show that the expansion feature of the abbreviation does a little help. We consider that the above situation may have a lower probability of appearing in the abbreviation dataset.

4.3.2 Comparison experiments results.

The result of the comparative experiment on two abbreviation recovery method is shown in Table 5.

From Table 5, we can see that the HMM-based abbreviation recovery method has similar performance to CRF on Top-all results, but is significantly worse in Top-1 and Top-2 than the proposed method in this paper. Since the abbreviations have a one-to-many mapping relationship, and the same abbreviations in different documents may express different meanings. While the HMM method does not consider the current context of the abbreviations for each word when the word is recovered. Thus, for abbreviations with different contexts, predicted result using the HMM-based abbreviation recovery method is always the similar. Experiments show that the proposed recovery algorithm based on improved CRF in this paper can obtain more stable results than the comparative test method in different situations.

Table 5. The result of the comparison between HMM method and CRF method

Model	Top-N	Precision
CRF	1	**0.621**
	2	**0.691**
	all	**0.934**
HMM	1	0.413
	2	0.525
	all	0.886

5 Conclusion

In order to solve the challenging problem of network supervision, the theme word substituted by abbreviation, this paper proposes a collaborative recognition and recovery method of intercept abbreviations. Firstly preprocess the texts by segmentation, and then further cut the incorrectly segmented texts, to get the candidate abbreviations. Secondly, use SVM to determine the candidate abbreviations with the devised statistic features. Thirdly, at recovery stage, use our improved CRF and statistic features of Chinese abbreviation to finally infer the corresponding originates. The experiments demonstrate that our model is effective and can get better results compared with other methods.

Our method can be used at other Chinese text content security analysis either. For example, improve the Chinese segmentation tool, automated update the vocabulary of the segmentation tool etc. In the future, we will try to improve the model to recognize and recover the abridged and concluded abbreviations.

References

1. Wang H F: Survey: Abbreviation Processing in Chinese Text. Journal of Chinese Information Processing 25(5), 60–67 (2011).
2. Wang A: Mining Informal Language from Chinese Microtext: Joint Word Recognition and Segmentation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 731–741. ACL, Sofia, Bulgaria (2013).
3. Wang A: Chinese Informal Word Normalization: an Experimental Study. In: The 6th International Joint Conference on Natural Language Processing (IJCNLP), pp. 127–135. ACL, Nagoya, Japan (2013).

4. Li C: Improving Named Entity Recognition in Tweets via Detecting Non-Standard Words. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, pp. 929–938. ACL, Beijing, China (2015).
5. Monroe W: Word Segmentation of Informal Arabic with Domain Adaptation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 206–211. ACL, Baltimore, Maryland, USA (2014).
6. Barrena A: Alleviating Poor Context with Background Knowledge for Named Entity Disambiguation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1903–1912. ACL, Berlin, Germany (2016).
7. Chang J S: A Preliminary Study on Probabilistic Models for Chinese Abbreviations. In: Proceedings of the 3rd SIGHAN workshop on Chinese language learning, pp. 9–16. ACL, Barcelona, Spain (2004).
8. Roark B: Hippocratic Abbreviation Expansion. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 364–369. ACL, Baltimore, Maryland, USA (2014).
9. Jiao Y: Abbreviation Prediction Using Conditional Random Field and Web Data. Journal of Chinese Information Processing 26(2), 62–68 (2012)
10. Zhang L K: Predicting Chinese Abbreviations with Minimum Semantic Unit and Global Constraints. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1405–1414. ACL, Doha, Qatar (2014).
11. Zhang L K: Coarse-grained Candidate Generation and Fine-grained Re-ranking for Chinese Abbreviation Prediction. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1881–1890. ACL, Doha, Qatar (2014).
12. Chen H: Chinese Named Entity Abbreviation Generation Using First-Order Logic. In: The 6th International Joint Conference on Natural Language Processing (IJCNLP), pp. 320–328. ACL, Nagoya, Japan (2013).
13. Shi Y Y: Cluster based Chinese Abbreviation Modeling. In: 15th Annual Conference of the International Speech Communication Association, pp. 273–277. COLIPS, Singapore (2014).
14. Chen F: Open Domain New Word Detection Using Condition Random Field Method. Ruan Jian Xue Bao/Journal of Software 24(5), 1051-1060 (2013).
15. Lavergne T: From n -gram-based to CRF-based Translation Models. In: Proceedings of the 6th Workshop on Statistical Machine Translation, pp. 542-553. ACL, Edinburgh, Scotland, UK (2011).
16. Tsuruoka Y: Stochastic gradient descent training for L1-regularized log-linear models with cumulative penalty. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pp. 477-485. AFNLP, Suntec, Singapore (2009).
17. Sokolovska N: Efficient Learning of Sparse Conditional Random Fields for Supervised Sequence Labeling. IEEE Journal of Selected Topics in Signal Processing 4(6), 953–964 (2010).
18. Yin Q: A Joint Model for Ellipsis Identification and Recovery. Journal of Computer Research and Development 52(11), 2460-2467 (2015).
19. Sun X: Learning Abbreviations from Chinese and English Terms by Modeling Non-Local Information. ACM Transactions on Asian Language Information Processing (TALIP) 12(2), 5:1-5:17 (2013).
20. Kenyon-Dean K: Verb Phrase Ellipsis Resolution Using Discriminative and Margin-Infused Algorithms. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1734–1743. ACL, Austin, Texas (2016).