On the Semi-unsupervised Construction of Auto-Keyphrases Corpus from Large-scale Chinese Automobile E-Commerce Reviews

Yang Li[1], Cheng Qian[1], Haoyang Che[2], Rui Wang[3], Zhichun Wang[1], Jiacai Zhang[1]

[1] College of Artificial Intelligence, Beijing Normal University, Beijing, China
[2] Data Intelligence Lab, Auto-smart Inc., Beijing, China
[3] Princeton International School of Mathematics and Science, Princeton NJ 08540, US
jiacai.zhang@bnu.edu.cn

Abstract. The long-standing automobile e-commerce websites in China have accumulated huge amounts of auto reviews, and extracting keyphrases of these reviews can assist researchers and practitioners in obtaining online users' typical opinions and acquiring their underlying motivations. However, there haven't existed any relevant text corpora so far. In this paper, the authors propose a semi-unsupervised scheme to construct a comprehensive auto-keyphrases corpus from online collected reviews in Chinese automobile e-commerce websites by Position Rank, which performs very well in keyphrases extraction from texts in the scenario of scarce labeled data. The iterative annotation process consists of three-round labeling and two-round corrections. During the process of the three-round unsupervised labeling, the computing model will extract seven most important words as the keyphrases of the whole paragraph. Between each labeling phase, there are manual check, correction, re-check and arbitration stages, in which the previous labeling errors are corrected and new vocabulary and rules are summarized up to further improve the unsupervised model. For comparison, the paper runs the experiments using another two unsupervised approaches: TF-IDF and Text Rank, the experimental results also show that Position Rank is a more efficient and effective method for keyphrases extraction. By the time this paper was written, the auto-keyphrases corpus had contained 110,023 entries, and there are still much room for improvement in corpus volume and labeling quality.

Keyphrases: Auto-Keyphrases Corpus, Keyphrases Corpus, Chinese Corpus, E-Commerce Website Reviews, Position Rank, Semi-unsupervised Method.

1 Introduction

1.1 Background

According to the survey data of the National Bureau of Statistics, the automobiles sales in China has reached 27.819 million in 2018, and declined 4.1% compared with the previous year [1]. The auto market in China is very huge, but now it has transferred from a high-speed growth stage to a low-speed and steady growth stage [2]. Inspired by new technologies such as artificial intelligence, the traditional automobile retail models need innovation and changes. To this end, automobile makers and retailers need to obtain a deep sense of users’ requirements quickly and accurately. One effective way to get end users' opinions is to retrieve and distill the user reviews on the auto internet websites. Over the past years, the Chinese automobile e-commerce websites (CAEW) have accumulated huge amounts of user reviews, based on which researchers and practitioners can utilize the up-to-date NLP technologies to gain an exact and instant understanding of auto users' underlying perspectives and the most authentic needs.

Thus, a large-scale and high-quality keyphrases corpus is required to lay the groundwork for further research and experiments. Recent years have seen very few text corpora in the auto field, and even fewer keyphrases corpora extracted from auto reviews. In a short term, to obtain enough labeled data is costly and time-consuming, so this paper devises to establish a semi-unsupervised approach incorporating Position Rank, an unsupervised method and manual efforts to construct the corpus, which widely ingest comments data from the most popular CAEW.

However, for the time being, existing methods relied on supervised methods with labeled data, but few of them adopted the unsupervised method in terms of scenarios with scarce labeled data. What we have constructed in the vertical field can be considered as a beneficial attempt towards the construction of Chinese corpora in an unsupervised method. Prior researchers often applied Kappa or F value to measure the consistency in a supervised construction method, while unsupervised methods were rarely proposed. In this paper, we measure the corpus accuracy with the help of practical manpower.

The labeled review keyphrases will help researchers and practitioners in the auto field access the key concerns of end users and achieve accurate marketing effects in the short term [3].

1.2 Characteristics of CAEW Reviews

Reviews of CAEW have their own characteristics. The Chinese vocabulary lacks morphological changes [4], and the automobile reviews are distinctly different from those of other fields. All of these characteristics indeed play an important role in the process of developing labeling specifications and selecting labeling methods or models. In summarization, there are six main characteristics of CAEW reviews.

Firstly, different online users have posted huge amounts of comments on different car models. Digging out the user concerns hidden behind these reviews can help automotive enterprises achieve accurate marketing effects.

Secondly, online user reviews update very fast. The number of online users in the most popular CAEW is massive, and reviews and discussions are generated frequently. In the process of constructing the review corpus, newly posted reviews need special consideration, to find a good method that can handle these data properly.

Thirdly, comments may include emoticons and internet jargons. Nowadays Internet users like to express their attitudes using emoticons or popular online jargons, which cannot be apt for processing by the old-fashioned model or the existing dictionary generally. This requires brand-new methods or Chinese dictionaries to deal with.

Fourth, the online comments are mostly directed at special car models. Perhaps the comment bodies may not mention the car model explicitly, but an implicit object does exist in default, of which phenomenon needs special focus when generating keyphrases from website reviews.

Fifth, there are some irrelevant comments in the raw review data, such as short comments, date info replies, reviews for location check-in and so on. Instead of providing enough information, such reviews should be deleted from raw prepared data during data preprocessing.

Sixth, labeled data is always lacking. Only a few websites proactively contain keyphrases codified by the website editors, but most majorities of websites only have raw comment contents without keyphrases, which makes it more difficult to use supervised method to construct the review corpus. However, as we all know, data labeling is a time-consuming and labor-intensive activity which may cost a lot of resource investments and require a very slow decision-making cycle.

1.3 Organization

The rest of this paper is organized as follows: Section 2 introduces the related work about the construction of CAEW corpus or knowledge base. Section 3 discusses the labeling rules. Section 4 introduces the process of corpus construction and how we use the unsupervised method. Section 5 reveals the labeling results and gives an analysis of the results. Section 6 gives the future work and concludes the paper.

2 Related work

In the 1990s, the academia began to research on the topics and subjects of the Chinese corpora. After years of development, the Chinese corpora have greatly improved in terms of volume, efficiency, and consistency, which cover a wide range of areas, including education, medicine, culture, engineering and so on. According to the survey results, the relevant Chinese corpora in industries are much less than expected, and reviews corpora of CAEW are currently not available.

This paper proposes a keyphrases corpus drilled from CAEW reviews for the very first time. Prior to this, Zheng et al. [5] proposed a hierarchical diagnostic system of car engines, through the combination of static knowledge base and dynamic knowledge base. Shanghai Translation Network [6] contains bilingual language materials for auto IT industry in Chinese and English, but lacks relevant data for CAEW reviews. Guo et al. [7] put forward to construct the car evaluation knowledge base based on the fuzzy concept, that is, combing with the partial sequence relationship between fuzzy theory and different concepts. They constructed a knowledge base about car evaluation, which is only used to indicate a relatively simple relationship instead of complex relationships between objects. Wang et al. [8] suggested that building a knowledge base of car evaluation based on fuzzy association rules can represent more complex relationships and can be used for knowledge reasoning in related fields. However, the topics and views put forward in the automobile evaluation haven't been deeply excavated. Feng et al. [9] proposed to build an opinion-based ontology base for passenger cars, using OWL and open source ontology editing tools from Stanford University. They used the following five aspects: the concept, relationship, concept level, non-classification relationship of concepts and axiom, to describe the automotive evaluation ontology and conduct deep mining from the users' emotion and views.

Overall, fewer corpus in the automotive-related sector exists to the public, and hardly any focusing on the CAEW reviews. In terms of the construction methods, researchers generally use labeled data, or combine manual labeling with tool labeling, while the unsupervised labeling method is rarely used at present.

3 Specification for Labeling

3.1 Raw Data Selection

This paper selects the word-of-mouth and forum contents from the most popular inland CAEW as the original data sources, which include BITA (bitauto.com), XCar (www.xcar.com.cn), ATHM (www.autohome.com.cn) and PCauto (www.pcauto.com.cn). The chosen principle is defined primarily as that reviews must come from the most popular CAEW at present with high popularity, justifiable credibility, and huge quantities of user comments. The original corpus contains not only comments, but also information such as the car model, the date of comment generation, and the date of comment publication. The contents of each car models are stored continuously when selected. Following that, it will delete comments with only emoticon or shorter words. The original corpus is stored in a JSON format. In general, the method has totally processed 134,741 word-of-mouth reviews at the outset.

3.2 Granularity of Labeling

The proposed corpus in this paper sets paragraphs as the labeling unit. To extract keyphrases, the method needs an exact understanding of what the statement expresses. It is necessary to grasp the contextual information as well as to understand the fine-grained semantics. Labeling keyphrases for paragraphs can be carried out from two points of views. The first one is a paragraph view, in which each user's entire comment does not split up, just label keyphrases directly for whole paragraph. The second one is a sentence view, in which the paragraph is cut into sentences, and then label keyphrases for each sentence, and finally integrated all phrases as the keyphrases of the whole paragraph.

At this stage, we intended to choose the first approach and regard paragraphs as a unit of research. Another method will be tried as planned in the follow-up work to make a further comparison and scheme upgrade.

3.3 Specification for Labeling

To construct a corpus, it is necessary to confirm the specification for labeling in advance. Specification is helpful to ensure the well-organized development of the labeling work and to ensure the quality of the labeling [10]. It also lays a good foundation for the expanding corpus and constructing comprehensive corpus.

Firstly, create a new property. Comments have been stored according to different attributes for different users, including user name, car model, review, etc. The model creates a new property named "keyphrases" to store keyphrases generated by the unsupervised methods. Secondly, control the number of keyphrases. When labeling according to the whole paragraph, the number of keyphrases is limited to 5-7. Because the reviews in this section is relatively long, we generally extract 7 words or phrases as the initial keyphrases. If we will label keyphrases based on a sentence mode later, each sentence will generate one or two important phrases, and then the model will integrate all words as the keyphrases for the entire paragraph. Thirdly, choose the Part of Speech. We extract notional words as keyphrases as far as possible. Notional words generally refer to nouns, pronouns etc., and these words contain more information. While function words generally refer to adjectives, adverbs, prepositions, conjunctions etc., which contain less information [11]. Notional words help a lot more than function words when mining users' opinions about passenger cars.

3.4 Process for Labeling

For obtaining the keyphrases from CAEW, we will perform the following processes sequentially. First, collect raw data. Choose data from the most popular CAEW, which include forum block name, car model name, reviews, publication date, and access date. Rich information can help analysis the semantics of keyphrases comprehensively, it also plays an important role in mining the typical opinion of users later. Secondly, data pre-processing. Cleaning up the raw data by deleting some bad data, which can ensure the quality of unsupervised labeling. Thirdly, unsupervised labeling and error correction. There are three rounds unsupervised labeling process. Between each two rounds, there is an error correction process to check labeling errors manually. If researchers find a labeling error, they will correct it and summarize the cause of the error, and then add the new principles to original dictionary and rule library. After updating, we will use the new dictionary and rule library to make unsupervised labeling again, until the third round of labeling is completed, as seen in Fig. 1. After comparing the common unsupervised methods: Position Rank, TF-IDF and Text Rank, finally we choose Position Rank as our method for its better results. The other two methods are set as control methods in our research.

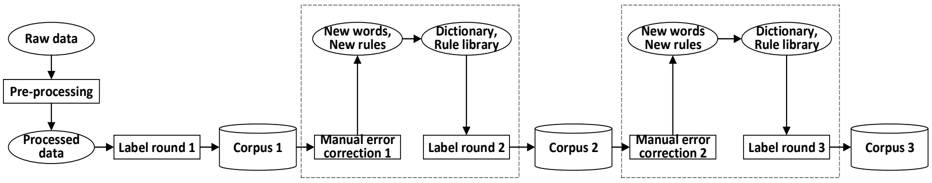

Fig. 1. The process of three-round unsupervised labeling and two-round error correction.

According to the requirements of error corrections, we team up with six other colleagues to complete all of the check, error correction, re-check, and arbitration tasks. Two of them are graduate students majoring in software engineering, responsible for checking and correcting errors. Another three are graduate students we recruited from the Faculty of Arts, responsible for the re-check work. The last one is a data scientist from an auto technology company, responsible for the final arbitration in terms of dispute. Before labeling, team members would conduct labeling training at first. We prepared 100 labeled reviews in CAEW in advance. All six team members labeled these data, and then checked the results and compared their labels with the original labels. When different opinions occurred, they would discuss and confirm which label was the best according to the related reviews, upon confirmation, they would modify the existing specification. The two-round error corrections lie somewhere in between the first round and second round of the unsupervised labeling. In this process, two students were responsible for checking and correcting the errors, respectively. After that, three students would re-check the labeled data. If the two students in charge of error corrections had different opinions about the labeled data, arbitration would be introduced at this time. The arbitration needs that all six colleagues organize a group discussion and judge which results are the best ones finally. At the same time, we would summarize some new vocabularies and extraction rules manually over those wrong items. Then we would add them into the original dictionary and extraction rules in the unsupervised method.

Through an iterative process, the labeling accuracy can be improved continuously. For the subsequent expansion of the reviews corpus, researchers can use these above-mentioned processes to ensure the consistency of the construction methods.

4 Construction of the Corpus

4.1 Data Pre-processing

Before labeling, we started with a cleaning of the 134,741 raw data, with an aim to improve the training efficiency and accuracy. And the following contents need to be deleted. At the beginning, the model should delete the irrelevant information from the comments, including the special symbols, emoticons, deactivated words, pictures, and so on. Secondly, the model should delete the comments that are too short and low-quality. Next, repetitive contents should be deleted too. Then, multiple comments with similar contents from the same user in the same post should be dealt with. Finally, unrelated comments should be deleted completely. After the deletion, 110,023 review items were available to the model.

4.2 Position Rank

Recently, keyphrases methods have already played a very important role in the inductive classification of text information and the theme search [12], this paper also focused on studying the construction method of the keyphrases corpus from the CAEW reviews. The research aims to extract the keyphrases from text, enrich the corpus contents, and facilitate the extensive utilization of the corpus as much as possible. The extraction of keyphrases is the critical step to the construct the keyphrases corpus. As of now, keyphrases extraction methods can be roughly divided into supervised learning methods and unsupervised learning methods [13]. In view of the CAEW reviews, such type of training sets makes supervised learning methods infeasible. Unsupervised learning methods are widely used in the emerging fields of keyphrases generation because they do not definitely require the training sets, and have developed rapidly [14] as a reliable and effective way of learning. This paper adopts a very popular unsupervised learning method based on the graph sorting (see Fig. 2).

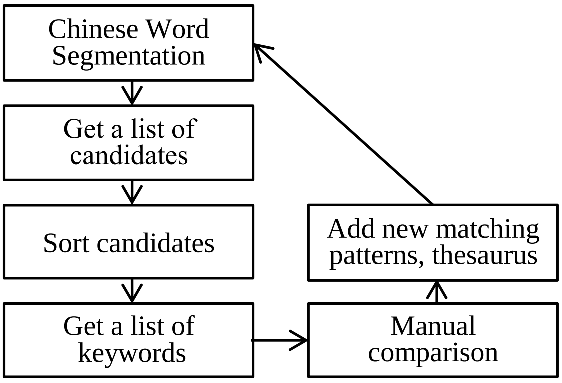

Fig. 2. Unsupervised learning method based on graph sorting, which uses the unlabeled data.

The unsupervised learning method based on graph sorting first needs to confirm the list of candidate words from text [15], out of the complexity of the Chinese language itself, this paper intends to confirm the list of candidate words by means of word labeling. Usually, keyphrases will be nouns or adjectives plus nouns form. However, the grams are more complex in Chinese, and because the comments will include a certain colloquialism, it is necessary to match more word-based labeling methods, such as verbs plus nouns and nouns plus adjectives [16]. When confirming the list of candidates, the model should consider as many lexical dimensions as possible.

Upon getting the list of candidates, the model starts building a graph based on them, where the nodes of the graph are the candidate words, and the model uses a fixed-size sliding window w to slide in the candidate list, if two candidates are in the same sliding window, they will be connected by a line between the nodes in the graph. Between the two candidates, a link exists. The graph can be directed or undirected, in order to facilitate the calculation of the fraction of the serrated nodes, and this paper adopts the undirected graph for computation.

The nodes in the undirected graph are candidates, and in order to find the keyphrases of text information, candidate words need to be evaluated, sorted from highest to lowest in terms of their importance. The evaluation criteria choose to use Position Rank [17], a PageRank [18] algorithm based on location bias. For candidate nodes in an undirected diagram, there are:

$$S(t+1) = \widetilde{M} \cdot S(t) \tag{1}$$

S in the upper formula represents the PageRank matrix, $\widetilde{M}$ represents the adjacent matrix of the undirected graph, and $S(t+1)$ is the PageRank matrix of $(t+1)$ time, multiplied by the adjacent matrix $\widetilde{M}$ and (t) time PageRank matrix.

Where the adjacent matrix $\widetilde{M}$ is subject to normalizing before calculation, the values in the matrix $\widetilde{m_{\iota\jmath}}$ are calculated as follows. V represents the undirected graph, $|V|$, which represents the number of nodes, is normalized to $\widetilde{m_{\iota\jmath}}$:

$$\widetilde{m_{\iota\jmath}} = \begin{cases} \widetilde{m_{\iota\jmath}} / \sum_{j=1}^{|V|} m_{ij}\,, \text{ if } \sum_{j=1}^{|V|} m_{ij} \neq 0 \\ 0, \text{ otherwise} \end{cases} \tag{2}$$

At the same time, in order to ensure that the undirected graph does not fall into the graph loop, the damping factor α is added, and the positional bias $\tilde{p}$ of the word is added, we get (3):

$$S = a \cdot \widetilde{M} \cdot S + (1-a) \cdot \tilde{p} \tag{3}$$

Where the position bias represents the position of candidate words in the text message, because the comments are the views expressed by the user, and according to the characteristics of the opinion selling, people tend to express their views at the beginning of the speech, so the candidate in the text message is more convincing than the candidate words. The specific calculation formula for the position bias $\tilde{p}$ is as follows:

$$\tilde{P} = \left[\frac{p1}{p1+p2+\cdots+p|V|}, \frac{p2}{p1+p2+\cdots+p|V|}, \dots, \frac{p|V|}{p1+p2+\cdots+p|V|}\right] \tag{4}$$

The initial score of the word, which is represented by p_1,p_2, is inversely proportional to the position of the word in the text, and the frequency of the words is proportional, if the first word appears in the article 5, 6, 7, then the $p_1 = \frac{1}{5} + \frac{1}{6} + \frac{1}{7}$, $p_1 + p_2 + \cdots + p_{|v|}$ Represents the total score of all words, and then divides the total score by p_1 to get the first word's share of all words, and finally gets the position bias $\tilde{p}$.

From these, we can get this:

$$S(v_u) = (1-a) \cdot \tilde{p} + a \cdot \sum_{v_j \in Adj(v_i)} \frac{w_{ji}}{O(v_j)} S(v_j) \tag{5}$$

The Position Rank score of the v_i node is identified by $S_{(v_i)}$ and α is the damping factor, and $\widetilde{p_\iota}$ is the positional bias of v_i. $S_{(v_j)}$ is the position Rank score of the v_j node, w_{ji} is the weight from the v_j node to the v_i node, $A\,dj(v_i)$ represents the accompanying matrix of the v_i node, and $O(v_j)$ represents all the out-bound weights of the v_j node.

Through the above Position Rank method, we can get a set of keyphrases for text information. The top 10 words can be included in the keyphrases list. Further, if the two combinations have appeared three or more times in the original text, the synthetic label is added to the keyphrases tag list, the value of Position Rank is the respective position value of the two keyphrases Rank. The resulting list of keyphrases is sorted from highest to lowest by Position Rank values, and the top five or the top seven keyphrases can be marked as keyphrases for comments. Through the processing of a large number of comments, a lot of labels are obtained, so as to build a keyphrases library based on the CAEW comments.

4.3 Process of constructing corpus

When the pre-processing is completed, the model needs to select a list of candidates in the comments. The Stanford NLP tool or other Chinese word breakers will be chosen for Chinese word segmentation, and this paper used the jieba word breaker. The matching word-sharing patterns are nouns, but the matching pattern is not fixed and unchangeable, according to the actual use the model can add a variety of matching patterns, such as various pronouns, verbs plus nouns or adjective forms, and so on. After the comments are manually labeled, we can observe the construction of the keyphrases syntax of the manual labels, increase or decrease the corresponding matching patterns, and improve the accuracy of the candidate words.

Once we get the list of candidates, we can implement the Position Rank algorithm. In the course of the experiment, this paper sets the damping factor α mentioned in the previous section to 0.85, and the number of keyphrases extracted is set to 7, wherein the sliding window w is set to 6, the iteration runs five times, and the comments and keyphrases list are combined to get the final keyphrases corpus.

The implementation method is not fixed and unchangeable, the damping factor α and sliding window w size in the algorithm parameters can be changed, the model selects the most appropriate parameter values according to manual evaluation results. At the same time, the model may not necessarily use the Position Rank algorithm or rigidly base on the figure-based sequencing of the unsupervised learning methods. The methods such as the Text Rank [19] [20] algorithm and the topic-based unsupervised learning method [21] [22] can be selected, however, this paper focuses on the idea of how to build a keyphrases corpus from CAEW comments, these alternative methods are not the focal points of this paper. The construction method aims only to gain enlightenment from academia and industry, and there are still many improvements left.

4.4 Quality Assurance

The method strictly controls the labeling quality through multiple rounds of unsupervised labeling and manual corrections. Firstly, it generates primary keyphrases labeled corpus (referred to as the primary corpus) through an unsupervised way. Secondly, it checks and corrects the constituent parts of the primary corpus. We randomly select 1000 comments from the primary corpus, and try to cover a variety of models and avoid the data gathered in a fixed certain mode. Because of the large-scale volume of the corpus, it is difficult to check all the data manually. We randomly retrieve 1000 entries more than once to detect the quality and improve the unsupervised model. Then we summarize the new vocabularies and extraction rules, and add them to the original dictionary and rule library. Thirdly, we use unsupervised method once again with the new dictionary and rule library to get the secondary corpus. Fourth, we randomly extract another 1000 entries, test the quality and summary rules. Finally, we use the unsupervised model once more, and then obtain the final corpus.

It is important to note that the data for the three rounds of unsupervised labeling are the same, and the data for the two random samplings are different. The labeling quality is guaranteed with the "three-round labeling and two-round correction" iteration.

5 Results and Analysis

5.1 Accuracy: Positon Rank

We list the accuracy in different rounds and changes between each two adjacent rounds using Position Rank. Besides the accuracy, we record the new words and new rules added to the old model. Table 1 gives the accuracy of Positon Rank in different rounds.

Table 1. Accuracy: Positon Rank.

Stage	Accuracy of the Position Rank (%)	Lift Rate(%)
1st-unsupervised labeling	33%	-
1st-artificial error correction	Add adjectives and verbs	-
2nd- unsupervised labeling	35%	+2%
2nd- artificial error correction	Add pronouns	-
3rd- unsupervised labeling	40%	+5%

We can see that the first round accuracy is 33%. At the first artificial error correction stage, we summarized adjectives and verbs are important factors for extracting, and added related words and rules into dictionary and rules library. The second round accuracy is 35%, and have an increase 2% from the first round. At the second artificial error correction stage, we add pronouns into rules library. The third round accuracy is 40%, and have an increase 5% from the second round, 7% from the first round. It reveals that iterative annotation is effective for unsupervised method.

5.2 Accuracy: Position Rank, TF- IDF, vs. Text Rank

In this section, we compared the accuracies of the three common unsupervised labeling methods: Position Rank, TF- IDF, and Text Rank. The three-round unsupervised labeling and two-round error corrections stays the same in all the three methods. The experimental results reveal that Position Rank obtain the best score after iterative annotation. Sometimes, the other two methods can get better keyphrases, which deserves in-depth research and discussion. Table 2 gives the accuracy rates of these different methods.

Table 2. Accuracy: Position Rank, TF- IDF, vs. Text Rank.

Stage	Positon Rank	TF-IDF	Text Rank
1st-unsupervised labeling	33%	34%	27%
2nd-unsupervised labeling	35%	37%	31%
3rd-unsupervised labeling	40%	35%	33%

6 Future Work and Conclusion

6.1 Future Work

There are several directions for further research and practices. First and foremost, the corpus needs the continuous expansion of the volume and quality. We will prepare several other medium-sized, niche CAEW as the original data sources, and process more forum blocks than the current method does. Secondly, the data preprocessing step needs further optimization. The website reviews may include many misspelled and misused words. If the error correction link is added in the pre-processing stage, we will further improve the labeling accuracy. Thirdly, the corpus asks for an infusion of the network jargons dictionary. The network jargons appeared with the development of the Internet in recent times and were used in high frequency by the net surfers. If the network jargons dictionary can be added into the unsupervised model, it can identify the users' comment semantics more effectively. Finally, it suggests that the corpus should be checked and corrected on the labeled data in a one-by-one basis manually to build a basic high-quality keyphrases corpus, but this may require a huge amount of manpower and time.

6.2 Conclusion

This paper proposes a semi-unsupervised scheme by the aid of Position Rank to construct the desired keyphrases corpus from CAEW reviews and get a relatively satisfactory result. The iterative annotation process consists of three-round unsupervised labeling and two-round manual error correction, which is proven to be an effective strategy. When correcting the error keyphrases, we summary new words and new rules which the original dictionary and rules library do not have. And we also add them into the unsupervised model to improve the accuracy.

At the same time, we tested another two methods: TF-IDF and Text Rank for comparison, and found that Position Rank is better at extracting keyphrases from users' reviews.

Acknowledgments
The work of this paper is funded by the National Key Technologies R&D Program (2017YFB1002502) and the project of Beijing Advanced Education Center for Future Education (BJAICFE2016IR-003). We would also like to thank Mr. Sheng Zhang for his help on correcting the paper and thank all of the teammates for their hard work.

References

1. National Bureau of Statistics Homepage, http://www.stats.gov.cn/tjsj/zxfb/201902/t20190228_1651265.html.
2. Shi Jianhua. What will happen to the low-growth car market? New Energy Vehicle News, 2019-05-13 (005).
3. Thousand City number zhi Guo Dengli: Opening up a new model for automobiles E-Commerce [J]. Internet Economy, 2019 (05): 102-103.

4. Yu Shiwen, Su Zhifang, Zhu Xuefeng. The Comprehensive Knowledge Base and Its Prospect.j.Chinese Journal of Informatics, 2011,25 (06): 12-20.
5. Xiaojun Z, Shuzi Y, Anfa Z, et al. A Knowledge-Based Diagnosis System for Automobile Engines[C]// Systems, Man, and Cybernetics, 1988. Proceedings of the 1988 IEEE International Conference on. IEEE, 1988.
6. Homepage, http://www.e-ging.com/article20150602064627/, last accessed 2019/6/12.
7. Guo Xiaomin, Wang Suge, Li Daewoo. Building a Knowledge Base for Automotive Evaluation based on Fuzzy Concepts//Proceedings of the CCIR2015.
8. WANG Su-ge,GUO Xiao-min,ZHANG Shao-xia.The Construction and Application of Automobile Evaluation Knowledge Based on Fuzzy Association Rules[J].Journal of Shanxi University(Natural Science),2016,39(03):423-428.
9. Feng Shufang,Wang Suge.The Construction of Automated Ontology Knowledge Base for Perspective Mining[J].Computer Applications and Software,2011,28(05):45-47+105.
10. Yu Shiwen,Zhu Xuefeng,Duan Huiming.The Processing Specification of Large-scale Modern Chinese Annotated Corpus[J].Journal of Chinese Information Processing,2000(06):58-64.
11. Khandelwal U, He H, Qi P, et al. Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context[J]. 2018.
12. Eibe Frank, Gordon W. Paynter, Ian H. Witten,Carl Gutwin, and Craig G. Nevill-Manning. 1999.Domain-specific keyphrase extraction. In Proceedings of the 16th International Joint Conference on Artificial Intelligence. pages 668–673.
13. Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 27th International Conference on Computational Linguistics. pages 1262–1273.
14. Kazi Saidul Hasan and Vincent Ng. 2010. Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In Proceedings of the 23rd International Conference on Computational Linguistics.pages 365–373.
15. Zhao JS, Zhu QM, Zhou GD, Zhang L. Review of research in automatic keyphrases extraction. Ruan Jian Xue Bao/Journal of Software, 2017,28(9):2431−2449 (in Chinese). http://www.jos.org.cn/1000-9825/5301.htm
16. Chang YC, Zhang YX, Wang H, Wan HY, Xiao CJ. Features oriented survey of state-of-the-art keyphrase extraction algorithms. Ruan Jian Xue Bao/Journal of Software, 2018,29(7):2046 2070 (in Chinese). http://www.jos.org.cn/1000-9825/5538.htm
17. Florescu C, Caragea C. A position-biased pagerank algorithm for keyphrase extraction. In: Proc. of the AAAI. Palo Alto: AAAIPress, 2017. 4923–4924.
18. Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine.Computer Networks, 30(1–7):107–117.
19. Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[C]// Proc Conference on Empirical Methods in Natural Language Processing. 2004.
20. Xiaojun Wan and Jianguo Xiao. 2008b. Single document keyphrase extraction using neighborhood knowledge. In Proceedings of the 23rd AAAI Conference on Artificial Intelligence, pages 855–860.
21. Grineva M P, Grinev M N, Lizorkin D. Extracting key terms from noisy and multitheme documents[C]// Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009. DBLP, 2009.
22. Liu Z , Li P , Zheng Y , et al. Clustering to Find Exemplar Terms for Keyphrase Extraction[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL. 2009.