Diagnosis Then Aggregation: An Adaptive Ensemble Strategy for Keyphrase Extraction

Xin Jin[1,2], Qi Liu[1,2(✉)], Linan Yue[1,2], Ye Liu[1,2], Lili Zhao[1,2], Weibo Gao[1,2], Zheng Gong[1,2], Kai Zhang[1,2], and Haoyang Bi[1,2]

[1] Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China, Hefei, China
[2] State Key Laboratory of Cognitive Intelligence, Hefei, China
{kingiv,lnyue,liuyer,liliz,weibogao,gz70229,sa517494, bhy0521}@mail.ustc.edu.cn, qiliuql@ustc.edu.cn

Abstract. Keyphrase extraction (KE) is a fundamental task in the information extraction, which has recently gained increasing attention. However, when facing text with complex structure or high noise, current individual keyphrase extraction methods fail to handle capturing multiple features and limit the performance of the keyphrase extraction. To solve that, ensemble learning methods are employed to achieve better performance. Unfortunately, traditional ensemble strategies rely only on the extraction performance (e.g., *Accuracy*) of each algorithm on the whole dataset for keyphrase extraction, and the aggregated weights are commonly fixed, lacking fine-grained considerations and adaptiveness to the data. To this end, in this paper, we propose an Adaptive Ensemble strategy for Keyphrase Extraction (AEKE) that can aggregate individual KE models adaptively. Specifically, we first obtain the multi-dimensional abilities of individual KE models by employing cognitive diagnosis methods. Then, based on the diagnostic abilities, we introduce an adaptive ensemble strategy to yield an accurate and reliable weight distribution for model aggregation when facing new data, and further apply it to improve keyphrase extraction in the model aggregation. Extensive experimental results on real-world datasets clearly validate the effectiveness of AEKE. Code is released at https://github.com/kingiv4/AEKE.

Keywords: Keyphrase Extraction · Ensemble Learning · Cognitive Diagnosis

1 Introduction

How to extract the needed information from the huge amount of unstructured knowledge is the fundamental problem in the field of natural language processing today [19,20,33]. Among the information extraction methods, keyphrase extraction (KE) has garnered significant attention [14,27,34] as it can enhance the efficiency of natural language processing and benefit numerous downstream tasks, such as information retrieval [14] and document summarization [25].

L. Fang et al. (Eds.): CICAI 2023, LNAI 14473, pp. 566–578, 2024.
https://doi.org/10.1007/978-981-99-8850-1_46

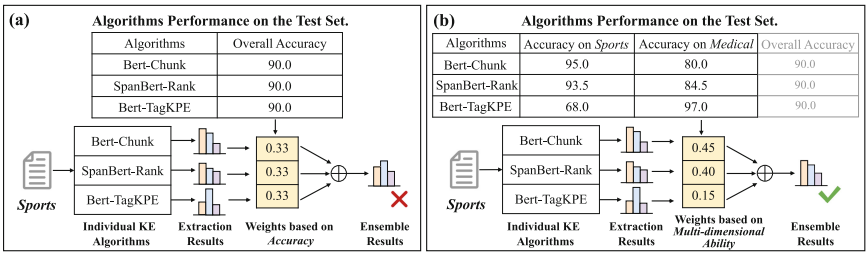

Fig. 1. Part (a) shows the traditional ensemble strategy based on *Accuracy*. Since the three methods perform consistently across the whole dataset, they are aggregated equally when encountering new data. However, part (b) shows that there are differences among the methods from a fine-grained perspective. When dealing with *Sports* news, more emphasis should be placed on the two methods (i.e. Bert-Chunk and SpanBert-Rank) that are more capable in *Sports*.

The goal of keyphrase extraction is to extract several keyphrases from documents that can represent the main information of the documents. For example, given a text document " *The authors had given a method for the construction of panoramic image mosaics with global and local alignment.* ", the keyphrase extraction method can identify "*panoramic image mosaics, global alignment, local alignment*" as the representative keyphrases. Finally, for evaluation, *Accuracy*, *Precision*, *Recall* and *F1-score* metrics are commonly employed to evaluate the performance of keyphrase extraction algorithms [10,29,32].

Despite previous approaches achieving promising results, when facing text with complex structure (e.g. long and difficult sentences) or high noise (e.g. text from different domains), these individual approaches fail to capture various features in the above text and have limited performance. To this end, a straightforward approach is to exploit the ensemble methods to aggregate different keyphrase extraction models to achieve better keyphrase extraction.

Figure 1(a) presents a traditional ensemble strategy that aggregates individual KE methods based on the *Accuracy*. However, unfortunately, in practice, the traditional methods can not always achieve satisfying results and even cause a negative impact on keyphrase extraction. Specifically, since SpanBert-Rank [28], Bert-Chunk, [28] and Bert-TagKPE [28] perform consistently in *Accuracy* on the overall dataset, we should aggregate these methods equally from the perspective of traditional ensemble strategy when facing new data about *Sports* topic. However, as shown in Fig. 1(b), from a more fine-grained perspective, both Bert-Chunk and Span-Rank outperform Bert-TagKPE on the *Sports* topic, while they perform poorly on the *Medical*. Therefore, when facing new *Sports* data, we should focus more on Bert-Chunk and SpanBert-Rank rather than dealing with all three methods equally during the model aggregation. In this paper, we define the extraction ability of the KE model for different topic domains as the multi-dimensional extraction ability of KE (See Sect. 3 for detail).

From the above observations, we can conclude that traditional ensemble methods fail to consider the multi-dimensional extraction abilities of individual models, and instead focus only on the performance of individual models

with a single metric (e.g., *Accuracy*), degrading the performance of the ensemble. Therefore, we argue that this "*ensemble pattern*" can be further explored to improve the keyphrase extraction.

Along this research line, in this paper, we propose an **A**daptive **E**nsemble strategy for **K**eyphrase **E**xtraction (AEKE) based on the multi-dimensional abilities of individual keyphrase extraction models. Specifically, inspired by the psychometric theories [5,22] from human measurement, we first diagnose the multi-dimensional abilities of different keyphrase extraction models by means of cognitive diagnostic techniques. Then, based on the diagnostic abilities, we develop an adaptive ensemble strategy. The strategy will adaptively adjust the aggregation weights for different samples to achieve better ensemble performance. Finally, experiments over two datasets, including OpenKP [32] and Inspec [15], validate the effectiveness of our AEKE.

2 Related Work

Keyphrase Extraction aims to select a set of phrases that could summarize the main topics discussed in the document [14]. The algorithms in keyphrase extraction are commonly divided into supervised and unsupervised methods. Specifically, unsupervised methods [2,3,24] mainly used different features of the document such as topic features, phrase frequency and so on to make keyphrase extraction. In supervised methods [7,29], pre-trained language models have been exploited and achieved competitive performance with annotation of the corpus.

Ensemble Learning can fuse the knowledge of individual models together to achieve competitive performance via voting schemes based on some learned features, which is widely used in machine learning tasks [8,25]. Traditional voting schemes include unweighted averaging and weighted voting. Among them, unweighted averaging of the outputs of the base learners in an ensemble is the most followed approach for fusing the outputs [11]. It considers the output results of each learner equally but ignores the differences between learners. On the other hand, weighted voting methods [11] tend to assign different weights to different learners based on their unidimensional ability. Such ability is often assessed by a single traditional metric on the history datasets. But the weights are constant during the model aggregation. In ensemble strategies of keyphrase extraction, mainstream methods employed unweighted averaging and weighted voting methods to aggregate individual KE models. However, these methods still suffered from relying on the unidimensional ability (e.g., *Accuracy*, *Precision*) of individual KE models to achieve aggregation, resulting in limited performance in the ensemble. To solve that, we develop an adaptive ensemble strategy for keyphrase extraction from the perspective of multi-dimensional abilities.

Cognitive Diagnosis is a fundamental task in many real-world scenarios (e.g., business [17] and education [12,13,31]). The main goal of cognitive diagnosis is to measure learners' proficiency profiles of abilities to finish specific tasks from their observed behaviors [31]. For instance in education, it can be used to infer student (as *learner*) knowledge proficiency (as *ability*) by fully exploiting

their responses of answering each exercise (as *task*). Most of the existing cognitive diagnosis models (CDMs) [5,12,22] are well designed from psychometric theories of human measurement. Among them, item response theory (IRT) [22] is the most classic CDMs which assumes the probability of the learner s_i correctly finishing a task e_j, i.e., $r_{ij} = 1$, increases with learner ability θ_i while decreasing with task difficulty β_j. Among them, the user ability and task difficulty are trainable unidimensional parameters [18]. A typical formulation of IRT is $P(r_{ij} = 1) = sigmoid((\theta_i - \beta_j) \cdot a_j)$, where a_j is an optional task discrimination item. Recently, some works extended the previous basic models to capture the more complex relationships among users, tasks, and abilities. The typical model is NeuralCD [31] which introduced neural networks $F(\cdot)$ to model high-level interaction between learners/abilities and tasks, i.e., $P(r_{ij} = 1) = F(\theta_i - \beta_j)$.

Inspired by the psychometric theories from human measurement, the multi-dimensional evaluation of KE algorithms can also benefit from the more fine-grained assessment of human learning performance.

3 Problem Definition

Cognitive Diagnosis for Keyphrase Extraction. Following the NeuralCD [31] which is a cognitive diagnostic model (CDMs), we introduce the definition of the cognitive diagnosis problem for keyphrase extraction algorithms. First, we denote the algorithms to be evaluated as learners and the CDMs as diagnosers. Then, with the diagnoser, we can evaluate the multi-dimensional abilities of learners on different skills, which are used to describe how well an algorithm performs on a particular category of samples.

Besides, in our work, since the topic of documents contains the main information and represents the specific textual features of keyphrase [23], we take the topics of documents as skills. For instance, topics on *Sports* and *Medical* convey a totally different message. Therefore, we define specific skills as specific topics of documents and one topic for one skill.

To design our diagnoser, we consider a well-trained learner set $S = \{s_1, ..., s_N\}$, a sample set $E = \{e_1, ..., e_M\}$ which is the dataset in our task, and a skill (topic) set $C = \{c_1, ..., c_P\}$. N and M denote the number of learners to be aggregated and samples in the dataset. P denotes the number of skills as a hyper-parameter in our task. Then the learner's output results on each sample as response logs R, which are denoted as a set of triplet (s, e, r_{ij}), where $s \in S$, $e \in E$ and r_{ij} is the score that learner i got on sample j. The top 5 results of keyphrase extraction are transferred to a score (0 or 1). We denote $r_{ij} = 1$ if learner i predicts more than one keyphrase correctly and $r_{ij} = 0$ otherwise. Meanwhile, an explicitly pre-defined sample-skill relevancy matrix Q should also be given. $Q = \{Q_{ij}\}_{M \times P}$, where $Q_{ij} = 1$ if sample e_i is related to skill p_j and $Q_{ij} = 0$ otherwise. Given the learner-sample response matrix R and the sample-skill matrix relevancy Q, we could estimate the multi-dimensional abilities of different learners on different skills through the diagnoser.

Adaptive Ensemble Strategy. Figure 1(a) illustrates the problems encountered with traditional ensemble strategies. They only focus on the performance of keyphrase extraction algorithms on a single metric, while ignoring the differences in multi-dimensional abilities. To solve that, from the perspective of the multi-dimensional abilities of the keyphrase extraction algorithms, we use the results of cognitive diagnosis to design adaptive ensemble strategies.

With the cognitive diagnostic module, we first obtain diagnostic results that include the multi-dimensional abilities of each algorithm and the characteristics (e.g., difficulty, discrimination, topic) of the data. Then, in the face of the new document n, we design the ensemble strategy of adaptive weight adjustment based on the above diagnostic results, including the multi-dimensional abilities, difficulty, discrimination, and topic. Among them, the multi-dimensional abilities represent the characteristics of the KE algorithms, while the difficulty, discrimination, and topic represent the characteristics of the samples. The goal of our strategy is to construct a relationship among diagnostic results and get more reasonable voting weights w for algorithms adaptively to get a better ensemble performance on every new document.

Problem Definition. *Given the multi-dimensional abilities of KE algorithms and features of the new document, our goal is to design an adaptive ensemble strategy to adjust the aggregation weights to improve the keyphrase extraction.*

4 Adaptive Ensemble Strategy via Cognitive Diagnosis

In this section, we present the details of AEKE for keyphrase extraction, which contains two stages. First, in the cognitive diagnostic stage, we follow NeuralCD [31] diagnostic approach and perform fine-grained diagnostics on the performance of various individual keyphrase extraction models to obtain their multi-dimensional abilities. In the ensemble stage, we design an adaptive ensemble strategy based on the diagnostic multi-dimensional abilities and document characteristics to get a better ensemble performance.

4.1 Cognitive Diagnose for Keyphrase Extraction Algorithms

Learner and Sample Factors. In our task, since we only focus on the ability of the different skills, each learner is represented with a one-hot vector $s_z \in \{0,1\}^{1\times N}$ as input, where N denotes the number of learners to be evaluated. In the same way, we represent sample e_d input as one-hot vector $e_d \in \{0,1\}^{1\times M}$.

Skill Factors. We want to make the topics as skills, as topic information is valuable in keyphrase extraction tasks. However, the published datasets do not contain topic labels for documents. To this end, in this paper, we employ the LDA [1] (Latent Dirichlet Allocation topic model) to obtain the topic labels by unsupervised clustering of the documents. Especially, LDA has better interpretability and the topical tokens for the clusters can be used as the explicit description for skills, which is great of importance for cognitive diagnosis.

After clustering the documents into P topics by LDA, we can obtain the sample-skill matrix $Q \in \{0,1\}^{M \times P}$. By this method, the topic label of each sample will be obtained.

Latent Factors. Following NeuralCD [31], with the model we can get the multi-dimensional abilities of the learner h_a and the difficulty h_d and discrimination h_d^{disc} of the sample. Among them, the h_a indicates the ability of the learner to process samples on different topics. The h_d represents the degree of difficulty the learner to solving the problem. Besides, the h_d^{disc} indicates the capability of samples to differentiate the proficiencies of learners. Samples with low discrimination mean that of low quality: they tend to have annotation errors or do not make sense.

Interaction and Prediction. Here, we exploit neural networks to model the relationship between learner ability factor h_a and skill difficulty factor h_d. The probability Y is defined as the ability compared with the sample in the covered topic as $Y = (h_a - h_d) \times h_d^{disc}$. Then, we use the full connection layers F to predict the score y of learner z on the sample d: $y = \sigma(F(Y))$. Finally, the whole objective of the diagnoser is defined with the cross entropy loss function:

$$\mathcal{L} = -\sum_i (r_i \log y_i + (1 - r_i) \log(1 - y_i)), \tag{1}$$

where r is the true score. Based on Eq. (1), we can get the multi-dimensional abilities of the keyphrase extraction algorithms.

4.2 Adaptive Ensemble Strategy

With the diagnostic module, we get the multi-dimensional abilities of each keyphrase extraction algorithm. Based on such diagnostic results, we propose an adaptive ensemble strategy to better aggregate the results of each extraction algorithm in the face of new test samples.

Inputs for Adaptive Ensemble Strategy. The inputs to the adaptive ensemble strategy include the abilities of individual KE algorithms and the features of the new sample. Among them, the KE algorithms' abilities are obtained from the previous diagnostic module, indicating the multi-dimensional abilities of the KE algorithms on different topics.

Features of the new sample contain information about the topic, difficulty and discrimination. Specifically, the topic information is associated with the diagnosed algorithm ability. The difficulty and discrimination information can reflect the implicit features of the algorithm in dealing with such problems to some extent. The information of the new sample is adequately represented by these three features.

To sum up, based on the topic model obtained in Sect. 4.1, the new sample is input and its distribution over each implicit topic is obtained as its topic information c_n. Each of its dimensions represents the probability of its distribution on that implicit topic.

Besides, since unseen samples are not used as input to the diagnostic module, the difficulty and discrimination of the samples cannot be directly obtained. To this end, we design a non-parametric module to predict the difficulty and discrimination. Specifically, as there is a close relationship between original texts and the factors of samples including the difficulty and discrimination, we choose to predict difficulty and discrimination based on semantic K-nearest neighbor [26] methods. Here, given the token sequence of original texts of keyphrase extraction samples $D^w = \{d_1^w, d_2^w, ..., d_n^w\}$, we map each word of D^w into word embedding by BERT [6], and get the document embedding by applying mean-pooling, where n_w is the length of the word sequence. We use the document embedding e_d as input representation for the new sample:

$$e_d = MeanPool(BERT([d_1^w, d_2^w, ..., d_{n_w}^w])). \tag{2}$$

Then, we match and retrieve the textual representations of the new samples with the representations of the samples entering in the diagnosis and find the K closest samples. These samples are able to get the corresponding difficulty $\{d_1, ..., d_k\}$ and discrimination $\{disc_1, ..., disc_k\}$ by diagnosis. Finally, we average the difficulty and discrimination retrieved as the difficulty d_n and discrimination $disc_n$ of the new sample.

Weight Prediction. After getting the above inputs, we need to get the most appropriate ensemble weights for each new sample. To ensure the interpretability of the weights, we design the ensemble strategy for the new samples by the following calculation:

$$w = SoftMax(h_a \cdot c_n \times d_n \times disc_n), \tag{3}$$

where $w \in \mathbb{R}^{N\times 1}$, $h_a \in \mathbb{R}^{N\times P}$, $c_n \in \mathbb{R}^{1\times P}$, d_n and $disc_n$ are single numbers.

5 Experiments

5.1 Experimental Setup

Dataset Description. We conduct experiments on two common keyphrase extraction datasets, i.e., OpenKP [32] and Inspec [15]. OpenKP is an open-domain keyphrase extraction dataset with various domains. In our settings, we choose the valid set (6,600 documents) of OpenKP for experiments. Besides, Inspec consists of short documents selected from scientific journal abstracts which are labeled by the authors, we choose the test (500 documents) and valid (1,500 documents) sets in this paper. The detailed statistics of the datasets are shown in Table 1. In particular, in our task, it is necessary to divide the dataset into two subsets, one for the diagnostician module and the other for the ensemble experiments. Therefore, we split the two datasets according to 3:1.

Algorithms to be Aggregated. To better train the diagnoser module and obtain the multi-dimensional abilities of each KE algorithm, we select 24 representative KE algorithms as follows:

- **Unsupervised methods:** Firstphrase[1], YAKE [4], TextRank [24], SingleRank [30], TopicRank [3], TopicalPageRank [21], PositionRank [9], MultipartiteRank [2], SIFRank [29].
- **Supervised methods:** BERT-RankKPE [28], SpanBERT-Variants*5 [28], BERT-ChunkKPE [28], BERT-SpanKPE [28], BERT-JointKPE [28], BERT-TagKPE [28], RoBERTa-Variants*5 [28].

Among them, supervised methods are trained on the OpenKP training set (134k documents). We obtain the response logs of learners on all samples on the datasets. Following the past research [31], we split the response logs into the training set, validation set and test set as 7:1:2.

Table 1. Statistics of keyphrase extraction datasets.

Statistics	OpenKP	Inspec
Document Number	6,616	2,000
Document Len Average	900	128
Keyphrase Average	2.2	9.8
Keyphrase Len Average	2.0	2.5

Table 2. Evaluation of all diagnosers through predicting learner performance on samples.

Methods	OpenKP			Inspec		
	AUC	Accuracy	RMSE	AUC	Accuracy	RMSE
DINA	0.563	0.545	0.559	0.538	0.512	0.578
IRT	0.576	0.540	0.542	0.560	0.545	0.544
NeuralCD	**0.914**	**0.869**	**0.340**	**0.883**	**0.762**	**0.379**

Baselines. For the cognitive diagnosis, we evaluate the performance of NeuralCD with other well-known CDMs (i.e., IRT [22] and DINA [5]). Among them, IRT is the most popular cognitive diagnosis method, it models students' latent traits and the parameters of exercises like difficulty and discrimination with a logistic-like function. DINA is the first method to design the Q-matrix and it uses binary variables to represent mastery of skills. NeuralCD [31] is a neural cognitive diagnostic framework, which leverages multi-layers for modeling the complex interactions of students and exercises, aiming to diagnose students' cognition by predicting the probability of the student answering the exercise correctly.

For the ensemble learning strategy, we choose to compare our approach with the average strategy and the weighted voting strategy. The weights are constant based on the performance of the history dataset evaluated on the traditional metrics (e.g., *shape Precision*). We also choose several individual keyphrase extraction methods from the 24 representative KE algorithms described before as baselines.

Experimental Settings. In our experiment, we use the pre-trained uncased BERT-based [6] model with 768 dimensions hidden representation as our tool. In our experiments, we set $P = 10$ for both two datasets. As the number of topics P is the most important hyper-parameter in AEKE, we conduct sensitivity experiments on it in Sect. 5.3. To set up the training process for the diagnostic module, we initialize all network parameters with Xavier initialization. The Adam optimizer [16] is used in the experiment while the learning rate is set to

[1] https://github.com/boudinfl/pke..

0.0002. We train all diagnosers for 20 epochs and select the best model on the validation set for testing. All experiments are run on two NVIDIA A100 GPUs.

5.2 Evaluation Metrics

Learner Performance Prediction. Generally, the ground truth of the ability of learners can't be obtained, it's difficult to evaluate the performance of cognitive diagnosis models. In most works, the prediction of learners' performance is an indirect way of evaluating the model. Evaluation metrics including *Accuracy*, *RMSE* (Root Mean Square Error), and *AUC* (Area Under the Curve) are chosen. Among them, better predictions have higher values in *Accuracy* and *AUC*, while the lower *RMSE* value, the better the prediction is achieved.

Model Aggregation. We realize model aggregation based on each learner's proficiency in the topics. The aggregation is tested on both OpenKP and Inspec datasets with several traditional keyphrase extraction metrics including *Precision*, *Recall*, and *F1-score*.

5.3 Experimental Results

Learner Performance Prediction. The experimental results are reported in Table 2, we have several observations as follows. First, NeuralCD performs the best on both OpenKP and Inspec, demonstrating NeuralCD can effectively evaluate the ability of keyphrase extraction algorithms. Besides, the traditional models including IRT and DINA perform poorly, which reflects that the relationship between learners' ability and samples' features is too difficult to capture, and

Table 3. Model aggregation results of popular keyphrase extraction models. The top part lists some unsupervised methods, the middle part lists the supervised methods, and the bottom part lists the ensemble methods.

Methods	OpenKP			Inspec		
	P@5	R@5	F_1@5	P@5	R@5	F_1@5
Firstphrase	19.5	36.7	23.6	24.0	15.0	17.3
YAKE [4]	12.1	29.1	16.7	21.0	13.6	15.5
TextRank [24]	5.5	14.2	7.9	31.7	19.2	22.6
SingleRank [30]	14.4	34.5	19.7	33.0	20.2	23.6
TopicRank [3]	14.4	30.3	19.6	28.2	16.9	20.0
BERT-JointKPE [28]	22.7	57.1	30.3	37.9	24.3	27.9
SpanBERT-RankKPE [28]	23.2	61.8	33.9	38.7	24.9	28.6
RoBERTa-TagKPE [28]	23.0	58.9	31.8	36.9	23.7	27.2
Averaging	23.7	61.0	33.5	39.1	25.0	28.9
Weighted Voting (*Precision*)	24.0	61.4	33.7	39.7	25.2	29.4
AEKE	**24.5**	**62.0**	**34.1**	**40.3**	**25.8**	**29.8**

indirectly proves the effectiveness of neural networks. Through NeuralCD, we can obtain highly reliable diagnostic results to be applied in the ensemble stage.

Model Aggregation. We compare our AEKE with the traditional aggregation methods (i.e., weighted voting and averaging) to illustrate the efficiency of our method as presented in Table 3. Among them, weights for weighted voting are obtained based on the overall performance (*precision*) of the history datasets of each keyphrase extraction algorithm. Such weights are constant during the model aggregation. In general, firstly, compared to supervised and unsupervised methods, both AEKE and the baseline ensemble strategy perform better than individual methods, demonstrating the necessity of ensemble. Besides, our adaptive ensemble strategy outperforms the ensemble baseline on both datasets, indicating the effectiveness of aggregation according to multi-dimensional abilities.

Hyper-Parameter Sensitivity Study. In our work, the number of skills P is a hyper-parameter, which determines how well the topics are clustered and also influences the design of the assessment skills. Therefore, in this section, we investigate the sensitivity of P. Figure 2 shows the performance of AEKE with different topic numbers P on the OpenKP dataset. The experiment shows a rising trend followed by a falling trend in the effectiveness of the ensemble result as the number of P increases. 10 is the best topic clustering number for the OpenKP. Specifically, when P is small, the result of document topic clustering is poor, which further affects the cognitive diagnosis of multi-dimensional abilities and the ensemble procedure. While $P > 10$, the ensemble results tend to be stable. Therefore, in this paper, we set P to 10 for our experiments.

Case Study. In this section, to further illustrate the effectiveness of AEKE, we show a high-quality sample in OpenKP and the ensemble results of weighted voting and AEKE in Fig. 3. Specifically, we aggregate the extraction results of three KE methods (i.e. BERT-Chunk [28], RoBERTa-Rank [28] and RoBERTa-Span [28]) by our strategy and traditional weighted voting strategy, respectively. In Fig. 3(a), we illustrate a detailed procedure of our AEKE. First, the new sample is entered into the diagnosis module and we can obtain the corresponding diagnosis results. It is obvious that this sample is a shooting report, which belongs to

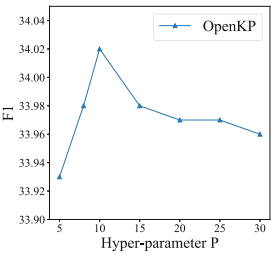

Fig. 2. Hyper-parameter Sensitivity Study.

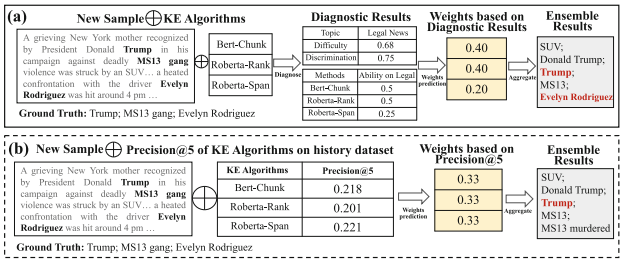

Fig. 3. Visualized keyphrases extracted by AEKE (a) and traditional strategy (b).

the legal news topic. Its difficulty and discrimination indicate that this sample has high text quality. It also shows the ability of the three methods on legal news topics. Then, based on the diagnosis results, our AEKE can adaptively adjust the weights of different methods during aggregation to get good ensemble results. In Fig. 3(b), the traditional method relies on the evaluation result of the three methods on the history datasets evaluated on the single metric *Precision@5*, and since the overall results on *Precision@5* are similar, the same weights are constant for all new samples. However, such weights do not achieve satisfying ensemble results in this new sample. This case serves as a compelling demonstration of the remarkable flexibility and efficiency of AEKE.

It is worth noting that, unlike traditional methods whose ensemble weights are fixed during aggregation, the weights in AEKE are not constant. Specifically, the above case belongs to the *Legal* topic, and when facing with samples of other topics (e.g., *Sports*), AEKE will adjust the ensemble weights adaptively based on the multi-dimensional abilities of KE methods and features of new sample.

6 Conclusion

In this paper, we proposed an adaptive ensemble strategy (AEKE) based on cognitive diagnostic techniques in the keyphrase extraction task. To the best of our knowledge, this is the first attempt to aggregate machine learning algorithms from a cognitive diagnostic perspective. To be specific, we first carefully employed the NeuralCD to evaluate the multi-dimensional abilities of keyphrase extraction algorithms. Then, based on the diagnostic ability, we developed an adaptive ensemble strategy to aggregate individual keyphrase extraction methods. Experimental results on both OpenKP and Inspec datasets demonstrated the effectiveness of AEKE.

Acknowledgements. This research was supported by grants from the National Key Research and Development Program of China (Grant No. 2021YFF0901003).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Boudin, F.: Unsupervised keyphrase extraction with multipartite graphs. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 667–672 (2018)
3. Bougouin, A., Boudin, F., Daille, B.: TopicRank: graph-based topic ranking for keyphrase extraction. In: International Joint Conference on Natural Language Processing (IJCNLP), pp. 543–551 (2013)
4. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A.M., Nunes, C., Jatowt, A.: A text feature based automatic keyword extraction method for single documents. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 684–691. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_63

5. De La Torre, J.: Dina model and parameter estimation: a didactic. Journal of educational and behavioral statistics **34**(1), 115–130 (2009)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Ding, H., Luo, X.: AttentionRank: unsupervised keyphrase extraction using self and cross attentions. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 1919–1928 (2021)
8. Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Front. Comp. Sci. **14**, 241–258 (2020)
9. Florescu, C., Caragea, C.: PositionRank: an unsupervised approach to keyphrase extraction from scholarly documents. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1105–1115 (2017)
10. Gallina, Y., Boudin, F., Daille, B.: Large-scale evaluation of keyphrase extraction models. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, pp. 271–278 (2020)
11. Ganaie, M.A., Hu, M., Malik, A., Tanveer, M., Suganthan, P.: Ensemble deep learning: a review. Eng. Appl. Artif. Intell. **115**, 105151 (2022)
12. Gao, W., et al.: RCD: relation map driven cognitive diagnosis for intelligent education systems. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 501–510 (2021)
13. Gao, W., et al.: Leveraging transferable knowledge concept graph embedding for cold-start cognitive diagnosis. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 983–992
14. Hasan, K.S., Ng, V.: Automatic keyphrase extraction: A survey of the state of the art. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1262–1273 (2014)
15. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pp. 216–223 (2003)
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Liu, C., Yang, L., Gao, W., Li, Y., Liu, Y.: MuST: an interpretable multidimensional strain theory model for corporate misreporting prediction. Electron. Commer. Res. Appl. **57**, 101225 (2023)
18. Liu, Q.: Towards a new generation of cognitive diagnosis. In: IJCAI, pp. 4961–4964 (2021)
19. Liu, Y., et al.: Technical phrase extraction for patent mining: a multi-level approach. In: 2020 IEEE International Conference on Data Mining (ICDM), pp. 1142–1147. IEEE (2020)
20. Liu, Y., et al.: TechPat: technical phrase extraction for patent mining. ACM Trans. Knowl. Disc. Data **17**, 1–31 (2023)
21. Liu, Z., Huang, W., Zheng, Y., Sun, M.: Automatic keyphrase extraction via topic decomposition. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 366–376 (2010)
22. Lord, F.: A Theory of Test Scores. Psychometric Monographs (1952)
23. Meng, R., Wang, T., Yuan, X., Zhou, Y., He, D.: General-to-specific transfer labeling for domain adaptable keyphrase generation. arXiv preprint arXiv:2208.09606 (2022)

24. Mihalcea, R., Tarau, P.: TextRank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. pp. 404–411 (2004)
25. Papagiannopoulou, E., Tsoumakas, G.: A review of keyphrase extraction. Wiley Interdisc. Rev. Data Min. Knowl. Disc. **10**(2), e1339 (2020)
26. Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
27. Song, M., Feng, Y., Jing, L.: A survey on recent advances in keyphrase extraction from pre-trained language models. In: Findings of the Association for Computational Linguistics, EACL 2023, pp. 2108–2119 (2023)
28. Sun, S., Liu, Z., Xiong, C., Liu, Z., Bao, J.: Capturing global informativeness in open domain keyphrase extraction. In: Wang, L., Feng, Y., Hong, Yu., He, R. (eds.) NLPCC 2021. LNCS (LNAI), vol. 13029, pp. 275–287. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88483-3_21
29. Sun, Y., Qiu, H., Zheng, Y., Wang, Z., Zhang, C.: SIFRank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model. IEEE Access **8**, 10896–10906 (2020)
30. Wan, X., Xiao, J.: Single document keyphrase extraction using neighborhood knowledge. In: AAAI, vol. 8, pp. 855–860 (2008)
31. Wang, F., et al.: Neural cognitive diagnosis for intelligent education systems. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 6153–6161 (2020)
32. Xiong, L., Hu, C., Xiong, C., Campos, D., Overwijk, A.: Open domain web keyphrase extraction beyond language modeling. In: Proceedings of the EMNLP-IJCNLP 2019, pp. 5175–5184 (2019)
33. Yue, L., Liu, Q., Du, Y., An, Y., Wang, L., Chen, E.: DARE: disentanglement-augmented rationale extraction. In: Advances in Neural Information Processing Systems (2022)
34. Zhao, H., Lu, M., Yao, A., Guo, Y., Chen, Y., Zhang, L.: Physics inspired optimization on semantic transfer features: an alternative method for room layout estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10–18 (2017)