Utilizing BERT Intermediate Layers for Unsupervised Keyphrase Extraction

Mingyang Song, Yi Feng and Liping Jing*
Beijing Key Lab of Traffic Data Analysis and Mining
Beijing Jiaotong University, China
mingyang.song@bjtu.edu.cn

Abstract

Extensive pre-trained language models such as the transformer-based BERT have been compelling at language modeling, achieving impressive results on numerous natural language downstream tasks. It has also been demonstrated that they implicitly retain factual knowledge in their parameters after pre-training. Understanding what the pre-training procedure of language models learns is critical to utilizing and enhancing them for Unsupervised Keyphrase Extraction (UKE). However, most existing BERT-based studies about UKE only use the single intermediate layer of BERT (e.g., the last layer) and ignore the latent knowledge in the intermediate layers. Therefore, in this paper, we analyze and explore the potential of utilizing BERT intermediate layers to enhance text representations and improve the performance of the state-of-the-art BERT-based unsupervised keyphrase extraction model. Specifically, we first verify and analyze the effect of adopting different BERT intermediate layers on the recent state-of-the-art unsupervised keyphrase extraction model. Then, based on the analysis, we propose a simple and effective feature aggregation strategy. Experimental results on several benchmark datasets demonstrate the effectiveness of aggregating intermediate layers of BERT to enhance text representations on the unsupervised keyphrase extraction task.

1 Introduction

Keyphrase extraction (KE) aims to extract a set of words or phrases from a given document that represents the salient information and main topics of the document (Hasan and Ng, 2014). KE models typically can be divided into supervised and unsupervised. Supervised approaches (Song et al., 2021) need large-scale annotated training data and perform poorly when transferred to different domain or type datasets. Unsupervised keyphrase extraction (UKE) approaches (Mihalcea and Tarau, 2004; Liang et al., 2021) are more universal and adaptive by extracting phrases based on information from the source document itself than the supervised method. This paper focuses on the unsupervised keyphrase extraction model.

A critical breakthrough in natural language processing is the use of heavily pre-trained transformers for natural language modeling, such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019). These Pre-trained Language Models (PLMs) are powerful for many downstream tasks in natural language processing and information retrieval, which have thus become an essential part in most cases. Therefore, research has also been done on BERT, especially to reveal what linguistic information is available in different parts of the model (Jawahar et al., 2019; de Vries et al., 2020). It has been noted that BERT progressively learns linguistic information roughly in the same order as the classic language processing pipeline: surface features are expressed in lower intermediate layers, syntactic features more in middle intermediate layers, and semantic ones in higher intermediate layers. Based on the above phenomenon, much recent work (Song et al., 2020, 2022) focuses on exploring the potential of utilizing BERT intermediate layers to enhance the fine-tuning performance of BERT. They demonstrated that each layer has different specializations, so combining information from different layers may be more beneficial instead of selecting a single one based on the best overall performance.

With the development of the pre-trained language models, recent unsupervised keyphrase extraction approaches (Sun et al., 2020; Liang et al., 2021; Ding and Luo, 2021) adopt the last layer of BERT as the embedding layer to obtain phrase and document representations instead of using the traditional pre-trained word vector, which significantly improves the performance of unsupervised

*Corresponding author.

keyphrase extraction. However, as mentioned earlier, the pre-trained language models (e.g., BERT (Devlin et al., 2019)) store rich language knowledge in the intermediate layers. Therefore, only using the single layer of BERT wastes the latent knowledge hidden in BERT. Meanwhile, judging a candidate phrase as a keyphrase also requires considering various features of natural languages (Hasan and Ng, 2014; Song et al., 2021), such as syntax, semantics, etc.

Motivated by the above phenomenon, we first probe the effectiveness of the intermediate layers of the pre-trained language model BERT on unsupervised keyphrase extraction. Then, we investigate several feature integration strategies for aggregating the middle layers of BERT to improve the performance of the state-of-the-art baseline (JoingGL (Liang et al., 2021)). Experimental results on DUC2001, Inspec, and SemEval2010 datasets show that combining intermediate layers of BERT as the embedding layer obtains better performance than using the single one.

2 Methodology

Given the sequence $\mathbf{x} = \{x_1, x_2, ..., x_m, ..., x_M\}$ with M tokens, we adopt BERT to encode it and obtain the hidden states for the i-th token from all L intermediate layers, $\mathbf{h}_i^L = \{h_i^1, h_i^2, ..., h_i^L\}$. Typically, L is set to 12. In this paper, based on the recent state-of-the-art model (JointGL (Liang et al., 2021)), we first probe the performance of different intermediate layers of BERT as the embedding layer on the current state-of-the-art model JointGL. Then, we test several feature integration strategies for combining the middle layers of BERT as the embedding layer on the baseline JointGL.

2.1 A Model-dependent Analysis

Existing embedding-based unsupervised keyphrase extraction models rely on the pre-trained language models (e.g., BERT) to achieve significant progress. Still, there is no work to probe in detail the impact of the features of different BERT intermediate layers on the unsupervised keyphrase extraction task. Therefore, to better utilize the pre-trained language models, we give a model-dependent analysis. Figure 1 shows the effectiveness of the baseline model JointGL by adopting different intermediate layers as the embedding layer on DUC2001, Inspec, and SemEval2010 datasets.

First of all, as can be seen from the results, the last layer of BERT is not always the one that gives the best performance of keyphrase extraction. Second, we found that the dependence on natural language features is different for different keyphrase extraction datasets when obtaining text representation. For DUC2001 and SemEval2001, the syntactic information in the middle intermediate layer is more critical. For Inspec, surface information is more critical. The above phenomenon is that the choice of features at different intermediate layers of BERT determines the effectiveness of phrases and documents representation. Overall, keyphrases in various datasets may rely on different linguistic features. Finally, an interesting phenomenon is that the 11-th intermediate layer achieves the worst F1, precision, and recall results. It is interesting to investigate why the 11-th layer obtains such results in future work.

2.2 Selective Feature Aggregation

We introduce a simple selective feature aggregation strategy to comprehensively use the linguistic knowledge stored in the pre-trained language models. Based on our model-dependent analysis, the intermediate layers with the higher evaluation metrics are selected as candidate layers in the first step. We generally choose the top K scores with their corresponding layers. For the second step, we use a weighted pooling operation to integrate the selected K layers as follows:

$$h_i^{'} = \gamma h_i^{k=1} + (1-\gamma) \sum_{k=2}^{K} h_i^k \qquad (1)$$

where γ is the balance factors and $h_i^{'}$ indicates the final representation of i-th input token. Through the above strategy, different feature information contained in different layers in BERT is fused to assist unsupervised keyphrase extraction.

3 Experiments

3.1 Datasets

To better verify the effectiveness of the proposed strategy, we evaluate our model on three benchmark keyphrase extraction datasets: DUC2001 (Wan and Xiao, 2008), Inspec (Hulth, 2003), and SemEval2010 (Kim et al., 2010).

3.2 Evaluation

We follow the common practice and evaluate the performance of our models in terms of f-measure

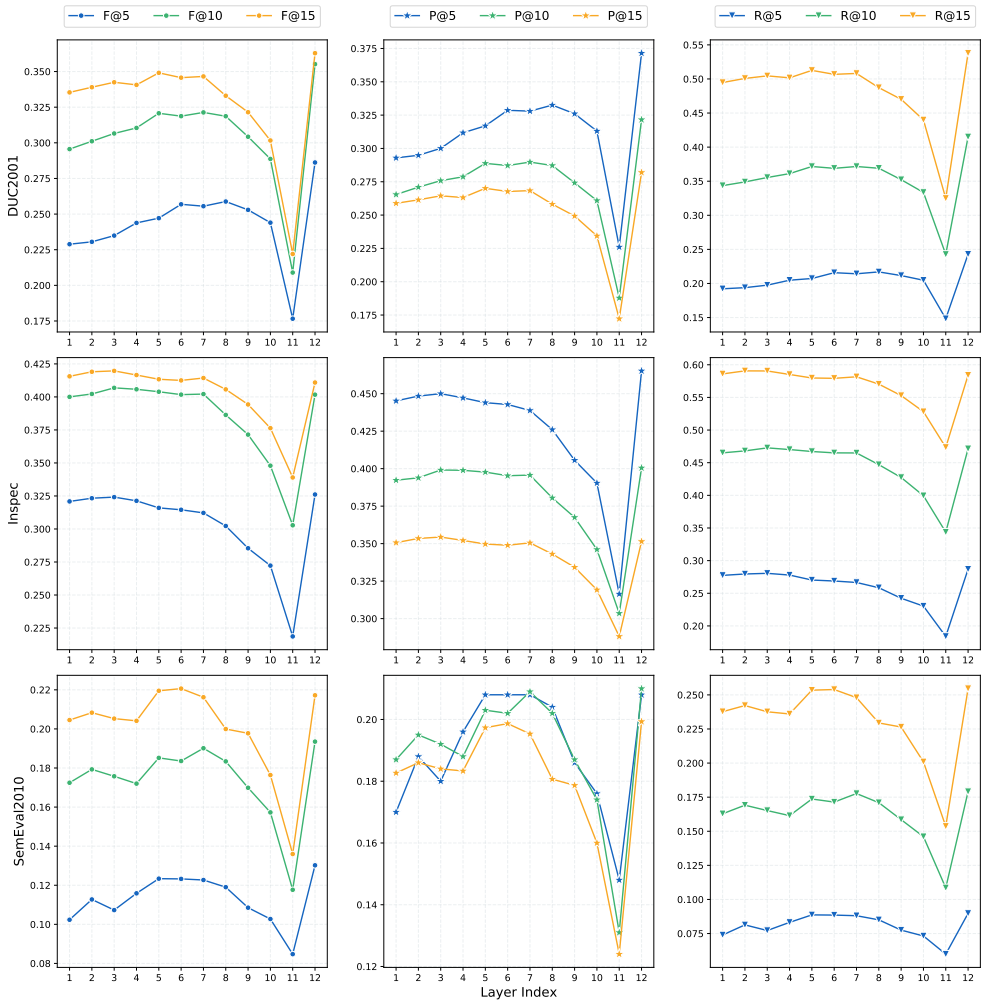

Figure 1: Results of the different intermediate layers of BERT for the baseline model JointGL on DUC2001, Inspec, and SemEval2010 test sets. The x-axis represents the index of the intermediate layers of BERT.

at the top N keyphrases (F1@N), and apply stemming to both extracted keyphrases and gold truth. Specifically, we report F1@5, F1@10 and F1@15 of each model on three benchmark keyphrase extraction datasets.

3.3 Implementation Details

We follow the previous baseline work (Liang et al., 2021) and adopt the same settings on the DUC2001, Inspec, and SemEval2010 datasets. For the selective layer aggregation strategy, γ is set to 0.95, 0.9, and 0.5 for DUC2001, Inspec, and SemEval2010. Specifically, K is set to 4, which means four intermediate layers with higher evaluation scores (F@5) are used in our strategy. Therefore, the $12, 8, 6, 7$-th intermediate layers are selected for DUC2001; the $12, 3, 2, 1$-th intermediate layers are selected for Inspec; the $12, 5, 6, 7$-th intermediate layers are selected for SemEval2010 (as shown in Table 1).

3.4 Results

Table 1 shows the results of F1, precision, recall, and @5, 10, and 15 using the embedding-based baselines and JointGL with different layer aggregation strategies on all datasets.

From the results in Table 1, we can see that the proposed feature aggregation strategy outperforms existing embedding-based unsupervised keyphrase extraction methods in most cases. The main reason is that the two most essential procedures of unsupervised keyphrase extraction methods are text representation and semantic similarity calculation. Our strategy obtains more comprehensive phrase and document representations by considering different linguistic knowledge stored in the pre-trained language models, which naturally improves the performance of the keyphrase extraction model.

Compared with different layer aggregation methods, our method has achieved significant improve-

Model	DUC2001			Inspec			SemEval2010		
	F1@5	F1@10	F1@15	F1@5	F1@10	F1@15	F1@5	F1@10	F1@15
Statistical Models									
TF-IDF (Jones, 2004)	9.21	10.63	11.06	11.28	13.88	13.83	2.81	3.48	3.91
YAKE (Campos et al., 2018)	12.27	14.37	14.76	18.08	19.62	20.11	11.76	14.4	15.19
Graph-based Models									
TextRank (Mihalcea and Tarau, 2004)	11.80	18.28	20.22	27.04	25.08	36.65	3.80	5.38	7.65
SingleRank (Wan and Xiao, 2008)	20.43	25.59	25.70	27.79	34.46	36.05	5.90	9.02	10.58
TopicRank (Bougouin et al., 2013)	21.56	23.12	20.87	25.38	28.46	29.49	12.12	12.90	13.54
PositionRank (Florescu and Caragea, 2017)	23.35	28.57	28.60	28.12	32.87	33.32	9.84	13.34	14.33
MultipartiteRank (Boudin, 2018)	23.20	25.00	25.24	25.96	29.57	30.85	12.13	13.79	14.92
Embedding-based Models									
EmbedRankd2v (Bennani-Smires et al., 2018)	24.02	28.12	28.82	31.51	37.94	37.96	3.02	5.08	7.23
EmbedRanks2v (Bennani-Smires et al., 2018)	27.16	31.85	31.52	29.88	37.09	38.40	5.40	8.91	10.06
SIFRank (Sun et al., 2020)	24.27	27.43	27.86	29.11	38.80	39.59	-	-	-
SIFRank+ (Sun et al., 2020)	30.88	33.37	32.24	28.49	36.77	38.82	-	-	-
KeyGames (Saxena et al., 2020)	24.42	28.28	29.77	32.12	40.48	40.94	11.93	14.35	14.62
AttentionRank (Ding and Luo, 2021)	-	-	-	24.45	32.15	34.49	11.39	15.12	16.66
MDERank (Zhang et al., 2021)	23.31	26.65	26.42	27.85	34.36	36.40	13.05	18.27	20.35
JointGL (Liang et al., 2021)	28.62	35.52	36.29	32.61	40.17	41.09	13.02	19.35	21.72
JointGL (Using the First Layers)	22.88	29.51	33.34	31.91	39.81	41.57	10.76	17.46	20.38
JointGL (Sum the 1-4 Layers)	23.60	30.44	33.78	32.06	40.33	41.64	11.63	17.73	20.28
JointGL (Sum the 5-8 Layers)	25.96	32.53	34.94	30.60	39.81	40.92	13.00	19.31	21.36
JointGL (Sum the 9-12 Layers)	24.31	28.26	29.94	27.25	34.71	37.56	10.66	15.64	17.69
JointGL (Sum the 1-12 Layers)	25.57	32.03	34.09	30.22	39.25	40.43	11.98	17.67	20.57
JointGL (Selective Feature Aggregation)	**28.92**	**35.71**	**36.54**	**32.40**	**40.71**	**41.92**	**13.25**	**19.93**	**22.23**

Table 1: Performance on DUC2001, Inspec and SemEval2010 test sets. F1 scores on the top 5, 10, and 15 keyphrases are reported. The best results of our model are bolded in the table.

ments, demonstrating the effectiveness of the proposed layer aggregation strategy.

Furthermore, it can be seen that different layer integration strategies have different effects on different datasets, which also shows the importance of potential knowledge mining in the pre-trained language model BERT.

4 Related Work

Unsupervised keyphrase extraction models mainly can be grouped into the traditional models (Jones, 2004; Mihalcea and Tarau, 2004; Bougouin et al., 2013) and embedding-based models (Sun et al., 2020; Bennani-Smires et al., 2018; Saxena et al., 2020; Liang et al., 2021). With the proposal and vigorous promotion of pre-trained language models, language models have become the backbone of most downstream natural language processing tasks, and significant progress has been made. Recent unsupervised keyphrase extraction models (Sun et al., 2020; Ding and Luo, 2021; Liang et al., 2021) adopt the last intermediate layer of the pre-trained language models as the embedding layer.

Different from the previous studies, this paper focuses on probing and aggregating the intermediate layers of the pre-trained language models for improving the performance of UKE.

5 Conclusions and Future Work

In this paper, we probe and analyze the effectiveness of aggregating the intermediate layers of BERT for unsupervised keyphrase extraction. To the best of our knowledge, we are the first work to probe BERT for unsupervised keyphrase extraction. Based on the findings of our non-parametric probing task, we propose a simple and effective feature integration strategy, which combines the intermediate layers of the pre-trained language models to improve the performance of UKE.

The main goal of this paper is to provide an empirical study of the existent models. Since we do not propose new models, there are no potential social risks to the best of our knowledge. Our work may benefit the research community by providing more introspection to the current state-of-the-art keyphrase extraction models. In future work, extending our work for self-supervised keyphrase extraction can also provide more insights into the utility of BERT for keyphrase extraction. Furthermore, it will be interesting to investigate why using the 11-th intermediate layer as the embedding layer leads the performance collapse.

References

Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. 2018. Simple unsupervised keyphrase extraction using sentence embeddings. In *CoNLL*, pages 221–229. Association for Computational Linguistics.

Florian Boudin. 2018. Unsupervised keyphrase extraction with multipartite graphs. In *NAACL-HLT (2)*, pages 667–672. Association for Computational Linguistics.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. Topicrank: Graph-based topic ranking for keyphrase extraction. In *IJCNLP*, pages 543–551. Asian Federation of Natural Language Processing / ACL.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. Yake! collection-independent automatic keyword extractor. In *ECIR*, volume 10772 of *Lecture Notes in Computer Science*, pages 806–810. Springer.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about bert's layers? a closer look at the nlp pipeline in monolingual and multilingual models. In *EMNLP (Findings)*, pages 4339–4350. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Haoran Ding and Xiao Luo. 2021. Attentionrank: Unsupervised keyphrase extraction using self and cross attentions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1919–1928.

Corina Florescu and Cornelia Caragea. 2017. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *ACL (1)*, pages 1105–1115. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *ACL (1)*, pages 1262–1273. The Association for Computer Linguistics.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *EMNLP*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL (1)*, pages 3651–3657. Association for Computational Linguistics.

Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, 60(5):493–502.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. Semeval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In *SemEval@ACL*, pages 21–26. The Association for Computer Linguistics.

Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2021. Unsupervised keyphrase extraction by jointly modeling local and global context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 155–164, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *CoRR*, abs/1907.11692.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *EMNLP*, pages 404–411. ACL.

Arnav Saxena, Mudit Mangal, and Goonjan Jain. 2020. Keygames: A game theoretic approach to automatic keyphrase extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2037–2048.

Mingyang Song, Yi Feng, and Liping Jing. 2022. Hyperbolic relevance matching for neural keyphrase extraction.

Mingyang Song, Liping Jing, and Lin Xiao. 2021. Importance Estimation from Multiple Perspectives for Keyphrase Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Youwei Song, Jiahai Wang, Zhiwei Liang, Zhiyue Liu, and Tao Jiang. 2020. Utilizing bert intermediate layers for aspect based sentiment analysis and natural language inference. *CoRR*, abs/2002.04815.

Yi Sun, Hangping Qiu, Yu Zheng, Zhongwei Wang, and Chaoran Zhang. 2020. Sifrank: A new baseline for unsupervised keyphrase extraction based on pre-trained language model. *IEEE Access*, 8:10896–10906.

Xiaojun Wan and Jianguo Xiao. 2008. Single document keyphrase extraction using neighborhood knowledge. In *AAAI*, pages 855–860. AAAI Press.

Linhan Zhang, Qian Chen, Wen Wang, Chong Deng, Shiliang Zhang, Bing Li, Wei Wang, and Xin Cao. 2021. Mderank: A masked document embedding rank approach for unsupervised keyphrase extraction. *CoRR*, abs/2110.06651.