Drop Redundant, Shrink Irrelevant: Selective Knowledge Injection for Language Pretraining

Ningyu Zhang[1,2 *], **Shumin Deng** [1,2 *], **Xu Cheng**[3], **Xi Chen**[5], **Yichi Zhang**[4], **Wei Zhang**[4], **Huajun Chen**[1,2 †]
[1] Zhejiang University & AZFT Joint Lab for Knowledge Engine
[2] Hangzhou Innovation Center, Zhejiang University
[3] National Engineering Laboratory for Improving the Government's Governance Capability Big Data Application Technology
[4] Alibaba Group
[5] Tencent
{zhangningyu,231sm,huajunsir}@zju.edu.cn
chengxu@pku.edu.cnm {yichi.zyc,lantu.zw}@alibaba-inc.com, jasonxchen@tencent.com

Abstract

Previous research has demonstrated the power of leveraging prior knowledge to improve the performance of deep models in natural language processing. However, traditional methods neglect the fact that redundant and irrelevant knowledge exists in external knowledge bases. In this study, we launched an in-depth empirical investigation into downstream tasks and found that knowledge-enhanced approaches do not always exhibit satisfactory improvements. To this end, we investigate the fundamental reasons for ineffective knowledge infusion and present selective injection for language pretraining, which constitutes a model-agnostic method and is readily pluggable into previous approaches. Experimental results on benchmark datasets demonstrate that our approach can enhance state-of-the-art knowledge injection methods.

1 Introduction

Self-supervised pre-trained language models (LMs) such as BERT, which can learn powerful contextualized representations, have achieved state-of-the-art results in natural language processing (NLP) tasks. However, open issues remain as these approaches lack domain-specific knowledge. Recent methods [Peters *et al.*, 2019] have revealed that the performance of the knowledge-driven downstream task (for example, question answering or relation extraction) is dependent on structured relational knowledge; thus, the direct finetuning of pre-trained LMs yields suboptimal results.

To address this issue, several works have attempted to integrate knowledge graphs (KGs) into pre-trained LMs [Zhang *et al.*, 2019; Levine *et al.*, 2020; Peters *et al.*, 2019; Xiong *et al.*, 2020; Zhang *et al.*, 2021a], which has shed light on promising directions for knowledge-driven tasks. Such methods generally retrieve pre-trained graph embeddings [Zhang *et al.*, 2019] or a KG subgraph via entity linking during pretraining and finetuning. Representations learned from knowledge-enhanced approaches have demonstrated expressive power and contributed to the performance improvement of downstream tasks. Thus, knowledge infusion has been widely adopted, as is a simple yet effective method that exploits external knowledge. Moreover, when sufficient training data are not available, the infusion of external knowledge into a pre-trained LM followed by finetuning to target tasks is more efficient [Zhang *et al.*, 2019].

To a certain extent, knowledge infusion integrates the knowledge which is insufficient into pre-trained representations and alleviates data requirements of the tasks. However, the adequate amount of external knowledge for effective infusion remains to be well understood. In recent years, [Petroni *et al.*, 2019; Broscheit, 2020] found that pre-trained LMs were partially equipped with a specific type of relational knowledge. Furthermore, [Liu *et al.*, 2020] observed that the incorporation of excessive knowledge might divert the context representation from its correct meaning. These observations motivated us to study the effective infusion of knowledge into pre-trained LMs. We note that previous approaches treated all external knowledge equally, thereby inevitably leading to redundant or irrelevant knowledge infusion. We argue that *knowledge is NOT always beneficial for downstream tasks, and an indiscriminate injection of knowledge may lead to* ***negative knowledge infusion****, which is detrimental to the performance of downstream tasks.*

In this paper, we take the first step towards studying this phenomenon fundamentally and propose general approaches to restraining detrimental knowledge during knowledge infusion.

Firstly, we investigate the efficacy of infused knowledge and observe that external knowledge (for example, entities) with high frequencies in the pre-trained corpus are more

[*] Equal contribution and shared co-first authorship.
[†] Corresponding author.

likely to trigger negative knowledge infusion. We argue that pre-trained LMs have already captured such external knowledge, and the redundant knowledge retrieved from KGs could possibly amplify the negative effects of the external noise, which subsequently deteriorates the performance. Inspired by this observation, we propose a **selective injection** mechanism that infuses informative knowledge by considering both the knowledge frequency and mutual reachability detected in the text for effective knowledge injection.

Secondly, we investigate those irrelevant parts of knowledge, which lead to the negative knowledge infusion regarding small spectral components. In particular, we conduct spectral analysis from the perspective of parameters, and feature representations based on singular value decomposition (SVD) [Golub and Reinsch, 2007] and make two observations: (1) small spectral components of weight parameters in high layers are not beneficial, and (2) when finetuning with sufficient training data, the small spectral singulars of the feature representations tend to **decay** autonomously. Inspired by these empirical observations, we leverage **spectral regularization** to suppress those small spectral components corresponding to irrelevant knowledge deliberately for effective *knowledge exploitation*. It should be noted that our approach is model-agnostic, and therefore orthogonal to existing approaches. We conduct numerous experiments on NLP benchmarks, which demonstrate the effectiveness in mitigating negative knowledge infusion. The contributions of this study can be summarized as follows:

- We investigate the problem of knowledge infusion into pre-trained LMs and observe that **redundant** and **irrelevant** knowledge exist for downstream tasks, which may lead to *negative knowledge infusion*.
- We then propose **selective injection** as well as **spectral regularization** respectively for effective knowledge infusion and our method is orthogonal to existing knowledge-driven tasks.
- Extensive experimental results on NLP benchmarks demonstrate the effectiveness of our method in alleviating negative knowledge infusion and our approach can enhance state-of-the-art knowledge injection methods.

2 Related Work

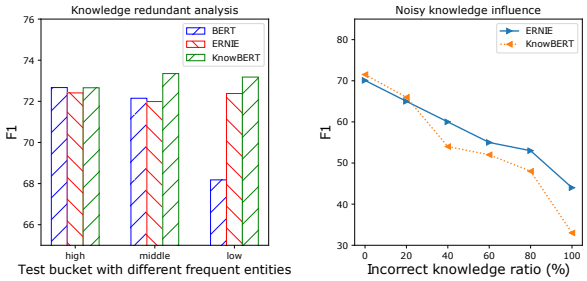
Background knowledge has been considered as an indispensable part of language understanding [Zhang *et al.*, 2021b], which has inspired knowledge-enhanced models including ERNIE (Tsinghua)[1] [Zhang *et al.*, 2019], ERNIE (Baidu) [Sun *et al.*, 2019], KnowBERT [Peters *et al.*, 2019], WKLM [Xiong *et al.*, 2020], LUKE [Yamada *et al.*, 2020], KEPLER [Wang *et al.*, 2019b], GLM [Shen *et al.*, 2020], K-Adaptor [Wang *et al.*, 2020], and CoLAKE [Sun *et al.*, 2020]. **ERNIE** [Zhang *et al.*, 2019] injects relational knowledge into the pre-trained model BERT, which aligns entities from Wikipedia to facts in WikiData. **KnowBERT** [Peters *et al.*, 2019] incorporates external KGs into BERT with a novel attention and re-contextualization approach. More recent methods, such as the GLM [Shen *et al.*, 2020], and K-Adapter [Wang *et al.*, 2020], introduce promising techniques to exploit informative knowledge and mitigate catastrophic forgetting during knowledge infusion. However, the dilemma of negative knowledge infusion is still not well understood.

Our work is motivated by approaches [Liu *et al.*, 2020; Petroni *et al.*, 2019; Broscheit, 2020] that have indicated the existence of redundant and irrelevant knowledge. Liu *et al.* [2020] observes that excessive knowledge incorporation could divert the context representation and Bian *et al.* [2021] finds that context-sensitive knowledge selection is critical, whereas [Petroni *et al.*, 2019; Broscheit, 2020] demonstrates that pre-trained LMs had been partially equipped with relational knowledge. Negative knowledge infusion, which is a largely ignored issue in recent knowledge-driven tasks, has rarely been considered. Moreover, our work is inspired by negative transfer in transfer learning [Chen *et al.*, 2019a] as they both follow a pretrain—finetune paradigm. However, as opposed to these approaches, we focus on knowledge infusion, including injecting favorable knowledge and exploiting beneficial representations. In contrast, transfer learning uses the knowledge acquired for one task to solve related tasks.

3 Knowledge-Enhanced Models

Regarding a knowledge-enhanced language model, when finetuning, it generally consists of two parts: a feature extractor F and a task-specific architecture C. We denote F^0 and C^0 as the pre-trained weights. We study the negative knowledge infusion, which is a phenomenon whereby the model infuses knowledge but does not achieve satisfactory improvement or even suffers from performance decay. It is natural to pose the following questions: 1) *Does negative knowledge infusion really exist in downstream tasks?* 2) *If it does, how does it affect the model performance?*

3.1 Negative Knowledge Infusion



(a) Redundant knowledge (b) Incorrect knowledge

Figure 1: Analysis of negative knowledge infusion. (a) F1 of subtest set with different-frequency entities; (b) influence of irrelevant knowledge.

In this section, we investigate whether negative knowledge infusion exists and whether it has a negative impact on task performance. We design two experiments based on ERNIE [Zhang *et al.*, 2019] and KnowBERT [Peters *et al.*, 2019] for

[1]In this paper, ERNIE refers to the ERNIE (Tsinghua).

evaluation[2].

In the first experiment, as illustrated in Figure 1(a), we sampled test samples and grouped them into buckets with different-frequency entities where the frequency refers to the occurrence number of entities in Wikipedia. Contrary to the common assumption, knowledge-enhanced approaches such as ERNIE and KnowBERT does not always exhibit satisfactory improvement to vanilla BERT and may even achieve slightly inferior performance with high-frequency entities for some instances, which indicates that not all knowledge is beneficial and demonstrates the existence of negative knowledge infusion. The pretrained language model may have already learned factual knowledge for high-frequent entities, which constitute redundant knowledge.

In the second experiment, as illustrated in Figure 1(b), we experiment to identify the influence of irrelevant knowledge by replacing the entity with other entities of different types. From Figure 1(b), we observe that irrelevant knowledge hurts the performance more severely as the noise rates increase. Note that there exist incorrect facts or wrong linked entities[3] which constitute irrelevant knowledge.

3.2 Why Negative Knowledge Infusion?

As negative knowledge infusion exists, we can ask another two questions: 1) *Which part of knowledge infusion causes negative knowledge infusion?* 2) *How can this problem be mitigated?*

From the perspective of knowledge, we begin to explore which part of the external knowledge may contribute to this problem. It can be observed from Figure 1(a) that there is no guarantee that the performance will always exhibit an improvement for samples with high-frequency entities. We argue that *redundant information may not contribute to the performance and irrelevant knowledge may hinder the performance*. Firstly, it should be noted that recent approaches [Petroni *et al.*, 2019; Broscheit, 2020] have demonstrated that pretraining can obtain relational knowledge. Since the pretrained LM has already captured such knowledge and several noisy facts exist in the external knowledge base, it is unreasonable to infuse this redundant external knowledge, resulting in noise and reducing the semantics in the text. Secondly, excessive knowledge may also lead to catastrophic forgetting, as observed by [Wang *et al.*, 2020].

From the perspective of features and parameters, we explore which part of the weight W and features $f = F(x)$ may not be beneficial. Figure 1(b) already illustrates that noises introduced either by incorrect facts or from erroneous entity linking may cause negative knowledge infusion. To further investigate the negative impact of irrelevant knowledge for downstream tasks, we analyze both the weights and features with principal angles [Rebuffi *et al.*, 2017], which have been introduced to measure the similarity of subspaces. Specifically, we use the corresponding angles [Chen *et al.*, 2019b],

[2]Negative knowledge infusion can also be found in Table 2.

[3]TagMe's performance on various benchmark datasets ranges from 0.37 to 0.72 F1 score [Kolitsas *et al.*, 2019]

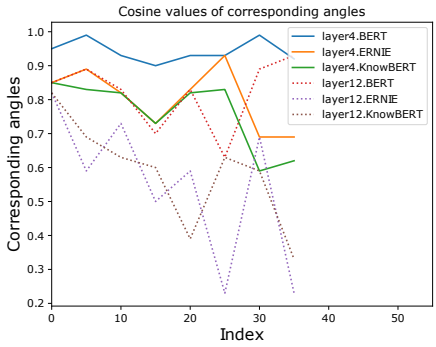


Figure 2: Cosine values of corresponding angles between W and W^0.

which are defined as follows:

$$\cos(\theta_i) = \frac{\langle \mathbf{u}_{1,i},\ \mathbf{u}_{2,i} \rangle}{\|\mathbf{u}_{1,i}\|\ \|\mathbf{u}_{2,i}\|} \tag{1}$$

where $\mathbf{u}_{1,i}$ refers to the i-th eigenvector with the i-th largest singular value and $\mathbf{u}_{2,i}$ denotes the opposite case. We apply θ to measure the availability of the eigenvectors in the weight matrices. Naturally, if the eigenvectors of the corresponding angle are small, the prior knowledge is more beneficial. Specifically, we denote $\mathbf{W}^0$ and $\mathbf{W}$ as the pre-trained weights of knowledge-enhanced models such as ERNIE and the fine-tuned weights on downstream tasks, respectively. We reshape the tensor into a matrix and subsequently perform SVD to obtain the eigenvectors $\mathbf{U}$ and singular values $\mathbf{\Sigma}$, denoted as follows:

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \tag{2}$$

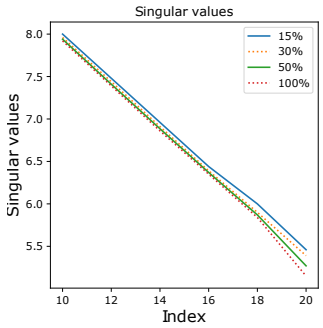


Figure 3: Singular values of feature matrices with different ratio of finetuning instances.

We calculate the relative angles θ in the 4-th layer (solid lines) and 12-th layer (dotted lines) between $\mathbf{W^0}$ and $\mathbf{W}$, as shown in Figure 2. We observe that the lower layers (4-th layer) have small relative angles, which is consistent with the finding in [Rogers *et al.*, 2020]. It is natural that lower layers are more beneficial for different tasks. Nevertheless, we note that relatively large singular values have rather small corresponding angles. Thus, it is intuitive to align weights

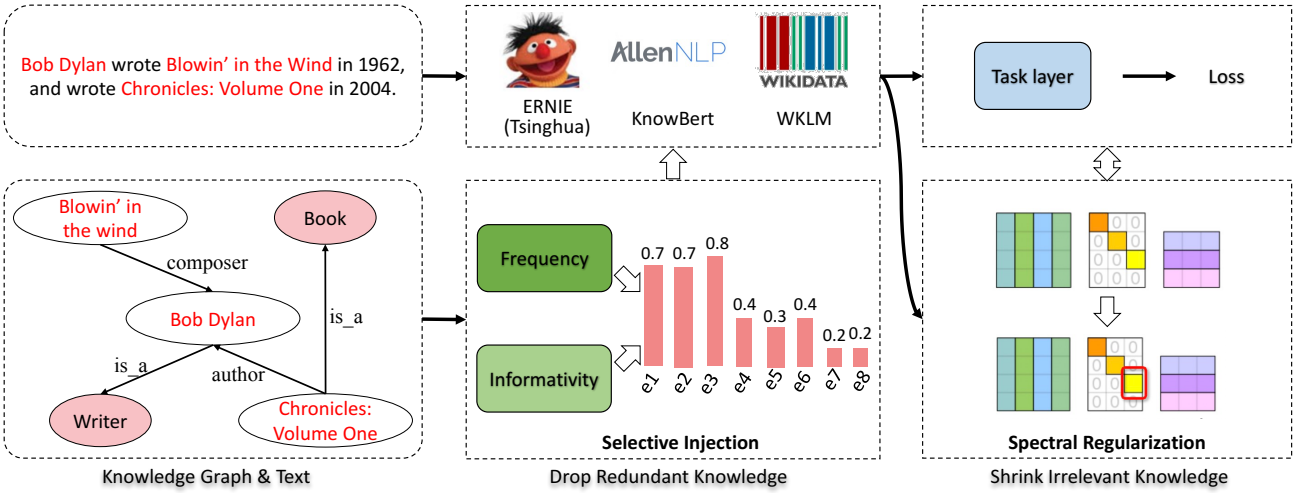


Figure 4: Knowledge infusion with **Selective Injection (SI)** and **Spectral Regularization (SR)**.

indiscriminately with the initial pre-trained values to remedy the negative knowledge infusion. Moreover, we analyze the feature representations with different training set sizes. Similarly, we use SVD to calculate all singular eigenvectors $\mathbf{U}$ and values $\mathbf{\Sigma}$ of the feature matrices, denoted as follows:

$$\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top} \tag{3}$$

As illustrated in Figure 3, we draw the diagonal elements of the singular value matrix $\mathbf{\Sigma}$ in descending order to measure the importance of the eigenvectors. As demonstrated in [Chen *et al.*, 2019a], finetuning and training from scratch can achieve comparable results with sufficient labeled data. It is natural to assume that finetuning with large datasets should provide greater generalization. Motivated by the observation of significantly suppressed, relatively small singular values of the features, we argue that promoting the similarity between these parts will give rise to negative knowledge infusion.

4 Approach

In this section, we preliminarily study how to alleviate negative knowledge infusion, as depicted in Figure 4. As the above analysis demonstrates that redundant knowledge is not necessary for infusion, it is intuitive to assign different sampling weights to different entities, thereby injecting different ratios of external knowledge. Moreover, as both weights and features with large singular values are valuable for downstream tasks, it is logical to shrink the importance of the lower spectral components with smaller scores, particularly with limited supervision. Note that the computation of SVD in high-dimensional weight spaces is costly; hence we mainly apply our approach on the feature space.

4.1 Dropping Redundant Knowledge with Selective Injection

The above redundant knowledge analysis of feature matrices results in the key inspiration. We propose a selective injection approach. Specifically, we randomly sample 85% of the injected entities as candidate knowledge and then introduce selective injection to infuse the necessary portions. Note that, as most frequently appearing entities are trivial and redundant, it is natural to assign lower sample probabilities to them. However, although several entities have a relatively high frequency, they cannot be neglected owing to their semantic importance. For example, one sampled entity should be assigned with a high probability if it can be inferred by numerous other entities in the same text (within $\mathrm{K}_{\mathrm{hop}}$-hops). To this end, we propose the selective injection approach regarding the following sampling equation [4]:

$$P\left(\mathcal{E}^{e_j}\right) \propto \mathbb{I}_{\{\mathrm{DF}(e_j)<\mathrm{K}_{\mathrm{thren}}\}} + \lambda\left[\left|\mathrm{S}\left(e_j\right)\right|\right]_{\mathrm{K}_{\min}}^{\mathrm{K}_{\max}}, \tag{4}$$

where $\mathrm{DF}(\cdot)$ refers to the document frequency, $\mathcal{E}$ is a set of linked entities from text, $\mathrm{S}(e) \triangleq \{e|\forall e' s.t.\,\mathrm{distance}\,(e', e) < \mathrm{K}_{\mathrm{hop}} \wedge e' \in \mathcal{E}\}$, $|\cdot|$ refers to the set size which denotes the number of neighbouring entities with distance shorter than K_{hop}, $[x]_a^b \triangleq \max(a, \min(x, b))$, and $\mathrm{distance}\,(e, e')$ is the shortest undirected length between the two entities. Note that the neglected knowledge in the selective injection still has 15% possibility to be infused into the LMs. Our approach can be used as a knowledge-sampling function for different knowledge-enhanced approaches.

4.2 Shrinking Irrelevant Knowledge with Spectral Regularization

Motivated by the above spectral analysis of the features, we propose a spectral regularization approach to remedy the irrelevant knowledge obtained during finetuning. In particular, we conduct SVD on the feature matrix F following Equation and penalize the smallest k singular values, as indicated below:

$$L_{\mathrm{sr}}(F) = \eta \sum_{i=1}^{k} \sigma_{i'}^2, \tag{5}$$

[4] λ and $\mathrm{K}_{\mathrm{hop/thresh/min/max}}$ are hyperparameters.

where k is the number of singular values to be penalized, η is a hyperparameter, and $\sigma_{i'}$ refers to the i-th smallest singular value.

Computational Complexity. The computational complexity of the selective injection can be ignored because it can be pre-computed prior to training. For a $a \times b$ matrix, the time complexity of the SVD is $O\left(\min\left(a^2b, ab^2\right)\right)$, which is unacceptable. We calculate the spectral regularization with $O\left(b^2d\right)$, where b is the batch size and d is the feature dimension (for example, 768). This is negligible in recent knowledge-enhanced approaches. Our approach can be embedded into existing fine-tuning scenarios, which can be formulated as:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} L\left(C\left(F\left(\mathbf{x}_i\right)\right), y_i\right) + \gamma\Omega(\mathbf{W}) + \eta L_{\mathrm{sr}}(F). \quad (6)$$

where L is the task loss, Ω is the L2 regularization, C is the task-specific function, L_{sr} is our spectral regularization, γ and η are hyperparameters.

5 Experiments

5.1 Datasets and Setup

TACRED [Zhang *et al.*, 2017] is a large-scale relation extraction dataset that covers 42 relation types and contains 106,264 sentences.

OpenEntity [Choi *et al.*, 2018] is a completely manually annotated entity typing dataset.

SearchQA [Dunn *et al.*, 2017] is a large-scale question answering dataset that is constructed to reflect a full pipeline of general question answering.

Quasa-T [Dhingra *et al.*, 2017] is a large-scale question-answering dataset consisting of 43,000 open-domain trivia questions and their answers that are obtained from various internet sources.

GLUE [Wang *et al.*, 2019a] is a benchmark with nine diverse NLP tasks. As WNLI mainly focuses on reasoning, we do not perform experiments on WNLI.

In this case, η was set to 0.001, k was set to 1, λ was set to 0.5, γ was set to 0.0001, $\mathrm{K}_{\mathrm{hop/thresh/min/max}}$ was set to {6,100,5,20}, and the batch size was set to 32. Note that our approach can also leverage other kind of external knowledge such as ConceptNet which is different from the world knowledge database Wikidata.

5.2 Baselines

We leverage Wikidata as an external knowledge base for both ERNIE and KnowBERT. We pretrain our own ERNIE and KnowBERT initialized with RoBERTa [Liu *et al.*, 2019]. We compare our approach with baselines as shown below:

BERT [Devlin *et al.*, 2018]. We utilize the BERT-base as the pre-trained language model baseline.

RoBERTa [Liu *et al.*, 2019]. We utilize the RoBERTa-base as baseline.

KEPLER [Wang *et al.*, 2019b]. It is a unified model for knowledge embedding and pre-trained language representation.

WKLM [Xiong *et al.*, 2020]. It is a weakly supervised pre-training approach that explicitly forces the model to incorporate knowledge about real-world entities.

K-Adaptor [Wang *et al.*, 2020]. It is an adaptor-based approach that fixed the pre-trained language model's parameters.

ERNIE* [Zhang *et al.*, 2019]. Here the model ERNIE* refers to the results obtained from the paper.

ERNIE. Here, the model ERNIE refers to our implementation results, which has the same amount of tuning with ERNIE+SI+SR.

KnowBERT* [Peters *et al.*, 2019]. Here the model KnowBERT* refers to the results obtained from the paper.

KnowBERT. Here, the model KnowBERT refers to the results of our implementation, which has the same amount of tuning with KnowBERT+SI+SR.

5.3 Results and Analysis

Main Results. From Table 1, we can observe the following: 1) ERNIE and KnowBERT embedded with our approach achieved improvement in all experiments and even performed better than RoBERTa in FIGER and TACRED, indicating the advantages of infusing informative knowledge and shrinking irrelevant features; 2) In SearchQA and Qusar-T, the improvement of our approach are relatively small, which could be owing to an insufficient quantity of available external knowledge, and thus fewer performance gains; 3) Both selective injection and spectral regularization contribute to the model performance, and selective injection obtains improvements in OpenEntity and TACRED, indicating the benefits of dropping redundant and irrelevant knowledge.

GLUE Results. From Table 2, we can observe the following: 1) ERNIE embedded with our approach achieved improvement in all experiments and obtained comparable results with RoBERTa-base on GLUE, further indicating the efficacy of our approach; 2) Our approach does not obtain much performance gains compared with RoBERTa-base. Note that those tasks are not knowledge-driven [Devlin *et al.*, 2018] which requires linguistic representations rather than structure facts; thus, knowledge-enhanced models such as ERNIE hurt the performance as it introduces noises, whereas our approach does not detour performances as it performs selective knowledge injection. It is advantageous for those indistinguishable situations whether knowledge is necessary or not (alleviating negative knowledge infusion).

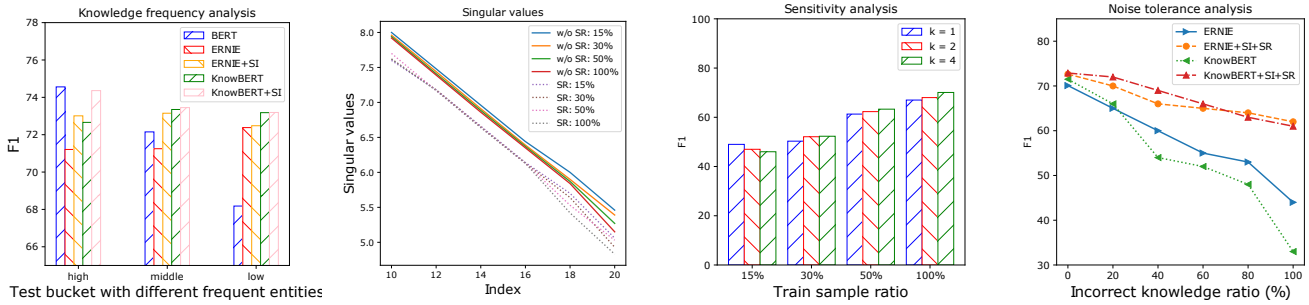
Selective Injection. To evaluate the effectiveness of the selective injection, we conducted ablation studies. It can be observed from Figure 5(a) that 1) the samples with high-frequency entities (redundant knowledge) exhibited a severe performance decay, which further demonstrates the negative impact of redundant knowledge; and 2) our approach with selective injection achieved more stable performance, suggesting that our mechanism of de-emphasizing the redundant knowledge was beneficial.

Model	OpenEntity			TACRED			SearchQA		Quasar-T	
	P	Ma-F1	Mi-F1	P	R	F1	EM	F1	EM	F1
BERT-base [Devlin *et al.*, 2018]	76.37	70.96	73.56	67.23	64.81	66.00	57.10	61.90	40.40	46.10
ERNIE* [Zhang *et al.*, 2019]	78.42	72.90	75.56	69.97	66.08	67.97	-	-	-	-
KnowBERT* [Peters *et al.*, 2019]	78.60	73.70	76.10	71.60	71.40	71.50	-	-	-	-
KEPLER [Wang *et al.*, 2019b]	77.20	74.20	75.70	70.43	73.02	71.70	-	-	-	-
WKLM [Xiong *et al.*, 2020]	-	-	-	-	-	-	61.70	66.70	45.80	52.20
RoBERTa [Liu *et al.*, 2019]	77.55	74.95	76.23	70.17	72.36	71.25	59.01	65.62	40.83	48.84
K-Adapter [Wang *et al.*, 2020]	79.25	75.00	77.06	70.05	73.92	71.93	61.96	67.31	45.69	52.84
ERNIE	78.52	72.92	75.62	70.92	69.28	70.09	59.53	65.92	44.35	51.15
ERNIE+SI	78.81	74.70	76.70	71.25	74.03	72.61	61.56	67.01	45.59	52.58
ERNIE+SI+SR	78.91	74.80	76.80	71.05	74.33	72.65	61.64	67.31	**45.79**	52.98
KnowBERT	78.63	73.80	76.14	71.50	71.50	71.50	60.93	65.92	44.45	50.95
KnowBERT+SI	78.61	74.73	76.62	71.15	73.73	72.42	62.66	67.32	45.70	52.88
KnowBERT+SI+SR	78.93	**75.56**	**77.21**	**71.35**	**74.49**	**72.89**	**62.86**	**67.52**	45.73	**53.10**

Table 1: Results on OpenEntity, TACRED, SearchQA, and Quasar-T datasets.

Model	MNLI	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	AVG.
RoBERTa	87.5/87.3	91.9	92.8	94.8	63.6	91.2	90.2	78.7	86.4
ERNIE	87.0/86.3	91.3	92.2	94.4	62.1	89.1	89.1	69.5	84.5
ERNIE+SI	87.1/87.0	91.5	92.0	94.4	62.2	90.0	89.3	75.6	85.4
ERNIE+SI+SR	87.4/87.1	92.0	92.3	94.6	63.3	90.6	90.5	76.5	86.0

Table 2: Results on different tasks of GLUE dev set.



(a) redundancy (b) σ (c) sensitivity analysis (d) noise tolerance

Figure 5: Analysis of selective infusion, spectral regularization, hyper-parameter sensitivity, and noise tolerance: (a) analysis of selective injection; (b) all singular values of feature matrices; (c) sensitivity analysis of different k; (d) analysis with different ratios of incorrect knowledge.

Spectral Regularization. The singular values of the features are drawn with (dotted) and without (solid) spectral regularization in Figure 5(b). We observed that 1) the singular values shrank, demonstrating the effectiveness of our approach; and 2) although $k = 1$ (k is the number of singular values to be penalized), more than one singular value was surprisingly suppressed, which shows the capability of the automatic distribution adjustment. We also conducted a sensitivity analysis of different k values using Equation 5. It can be observed from Figure 5(c) that 1) the performance of the limited training data with a larger k value was slightly superior; and 2) the performance of the sufficient training data decayed with a relatively large k, indicating the necessity of a trade-off between penalization and knowledge transfer.

Noise Tolerance. To further evaluate our approach's noise tolerance, we deliberately replaced entities with other entities of different types to simulate noisy facts in the knowledge base. We experimented with different ratios of noise in knowledge. According to 5(d), 1) all approaches exhibited a performance decay, indicating the negative effect of the irrelevant knowledge; and 2) our method significantly outperformed all of the baselines, suggesting that our approach was

more robust and could remedy the noisy effect resulting in negative knowledge infusion.

6 Conclusions and Future Work

We have studied the knowledge infusion of knowledge-driven tasks and took the first step towards delving into knowledge infusion scenarios from a new perspective: negative knowledge infusion. Whereas recent approaches have generally focused on designing sophisticated architectures to infuse knowledge, the essential mechanism of knowledge infusion remains less understood. We empirically observed two main findings, namely that *redundant and irrelevant knowledge will lead to negative infusion*, which may shed light on future works on knowledge-enhanced approaches. We proposed selective injection and spectral regularization to inhibit negative components, which can be embedded into existing methods demonstrated performance gains. We anticipate further research on promising directions, including 1) exploiting more efficient approaches to identify the useful knowledge; 2) investigating the essence of knowledge-driven tasks and proposing more effective infusion across domains.

Acknowledgments

We want to express gratitude to the anonymous reviewers for their hard work and kind comments. This work is funded by NSFCU19B2027/91846204, National Key R&D Program of China (Funding No.SQ2018YFC000004).

References

[Bian *et al.*, 2021] Ning Bian, Xianpei Han, Bo Chen, and Le Sun. Benchmarking knowledge-enhanced commonsense question answering via knowledge-to-text transformation. In *AAAI*, 2021.

[Broscheit, 2020] Samuel Broscheit. Investigating entity knowledge in bert with simple neural end-to-end entity linking. In *CoNLL*, 2020.

[Chen *et al.*, 2019a] Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In *NeurIPS*, pages 1906–1916, 2019.

[Chen *et al.*, 2019b] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*, pages 1081–1090, 2019.

[Choi *et al.*, 2018] Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. Ultra-fine entity typing. In *ACL*, 2018.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2018.

[Dhingra *et al.*, 2017] Bhuwan Dhingra, Kathryn Mazaitis, and William W Cohen. Quasar: Datasets for question answering by search and reading. *arXiv preprint arXiv:1707.03904*, 2017.

[Dunn *et al.*, 2017] Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.

[Golub and Reinsch, 2007] Gene H. Golub and Christian H. Reinsch. Singular value decomposition and least squares solutions. In *Milestones in Matrix Computation*, pages 160–180. Oxford University Press, 2007.

[Kolitsas *et al.*, 2019] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. End-to-end neural entity linking. In *CoNLL*, 2019.

[Levine *et al.*, 2020] Yoav Levine, Barak Lenz, Or Dagan, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. Sensebert: Driving some sense into bert. In *ACL*, 2020.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Liu *et al.*, 2020] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In *AAAI*, 2020.

[Peters *et al.*, 2019] Matthew E Peters, Mark Neumann, IV Logan, L Robert, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. Knowledge enhanced contextual word representations. In *EMNLP*, pages 43–54, 2019.

[Petroni *et al.*, 2019] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language Models as Knowledge Bases? In *EMNLP*, pages 2463–2473, 2019.

[Rebuffi *et al.*, 2017] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *NeurIPS*, pages 506–516, 2017.

[Rogers *et al.*, 2020] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. In *TACL*, 2020.

[Shen *et al.*, 2020] Tao Shen, Yi Mao, Pengcheng He, Guodong Long, Adam Trischler, and Weizhu Chen. Exploiting structured knowledge in text via graph-guided representation learning. In *EMNLP*, 2020.

[Sun *et al.*, 2019] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. In *ACL*, 2019.

[Sun *et al.*, 2020] Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. Colake: Contextualized language and knowledge embedding. In *COLING*, 2020.

[Wang *et al.*, 2019a] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, 2019.

[Wang *et al.*, 2019b] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. In *TACL*, 2019.

[Wang *et al.*, 2020] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*, 2020.

[Xiong *et al.*, 2020] Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model. In *ICLR*, 2020.

[Yamada *et al.*, 2020] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. Luke: deep contextualized entity representations with entity-aware self-attention. In *EMNLP*, 2020.

[Zhang *et al.*, 2017] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. Position-aware attention and supervised data improve slot filling. In *EMNLP*, pages 35–45, 2017.

[Zhang *et al.*, 2019] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced Language Representation with Informative Entities. In *ACL*, pages 1441–1451, 2019.

[Zhang *et al.*, 2021a] Ningyu Zhang, Qianghuai Jia, Shumin Deng, Xiang Chen, Hongbin Ye, Hui Chen, Huaixiao Tou, Gang Huang, Zhao Wang, Nengwei Hua, and Huajun Chen. Alicg: Fine-grained and evolvable conceptual graph construction for semantic search at alibaba. In *KDD*, 2021.

[Zhang *et al.*, 2021b] Shengyu Zhang, Dong Yao, Zhou Zhao, Tat-Seng Chua, and Fei Wu. Causerec: Counterfactual user sequence synthesis for sequential recommendation. In *SIGIR*, 2021.