ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation

Yu Sun[*]   **Shuohuan Wang**[*]   **Shikun Feng**[*]   **Siyu Ding**   **Chao Pang**

Junyuan Shang   **Jiaxiang Liu**   **Xuyi Chen**   **Yanbin Zhao**   **Yuxiang Lu**   **Weixin Liu**

Zhihua Wu   **Weibao Gong**   **Jianzhong Liang**   **Zhizhou Shang**   **Peng Sun**

Wei Liu   **Xuan Ouyang**   **Dianhai Yu**   **Hao Tian**   **Hua Wu**   **Haifeng Wang**

Baidu Inc.

{sunyu02, wangshuohuan, fengshikun01}@baidu.com

ABSTRACT

Pre-trained models have achieved state-of-the-art results in various Natural Language Processing (NLP) tasks. Recent works such as T5 [1] and GPT-3 [2] have shown that scaling up pre-trained language models can improve their generalization abilities. Particularly, the GPT-3 model with 175 billion parameters shows its strong task-agnostic zero-shot/few-shot learning capabilities. Despite their success, these large-scale models are trained on plain texts without introducing knowledge such as linguistic knowledge and world knowledge. In addition, most large-scale models are trained in an auto-regressive way. As a result, this kind of traditional fine-tuning approach demonstrates relatively weak performance when solving downstream language understanding tasks. In order to solve the above problems, we propose a unified framework named ERNIE 3.0 for pre-training large-scale knowledge enhanced models. It fuses auto-regressive network and auto-encoding network, so that the trained model can be easily tailored for both natural language understanding and generation tasks with zero-shot learning, few-shot learning or fine-tuning. We trained the model with 10 billion parameters on a 4TB corpus consisting of plain texts and a large-scale knowledge graph. Empirical results show that the model outperforms the state-of-the-art models on 54 Chinese NLP tasks, and its English version achieves the first place on the SuperGLUE [3] benchmark (July 3, 2021), surpassing the human performance by +0.8% (90.6% vs. 89.8%).

1 Introduction

Pre-trained language models such as ELMo [4], GPT [5], BERT [6], and ERNIE [7] have proved to be effective for improving the performances of various natural language processing tasks including sentiment classification [8], natural language inference [9], text summarization [10], named entity recognition [11] and so on. In general, pre-trained language models are learned on a large amount of text data in a self-supervised manner, and then fine-turned on downstream tasks or directly deployed through zero/few-shot learning without task-specific fine-tuning. Such pre-trained language models have become the new paradigm for natural language processing tasks.

In the past year or two, one of the important trends of pre-trained language models is their increasing model size, which leads to lower perplexity in pre-training and better performances on downstream tasks. Megatron-LM [12], with one billion parameters, is proposed for language understanding using a simple but efficient intra-layer model parallel

---

[*]Equal Contribution

approach, which achieves the state-of-the-art results on several datasets. T5 [1] explores the limits of pre-trained models with 10 billion parameters, but soon the record was broken by the GPT-3 model [2] with 175 billion parameters which has a good performance under the few-shot or even zero-shot settings. Soon afterwards, Switch-Transformer [13] is proposed as the world's first trillion-parameter pre-trained language model.

However, these large-scale pre-trained language models with hundreds of billions of parameters are trained on plain texts. For example, the 175-billion-parameter GPT-3 is trained on a corpus with 570GB filtered texts from Common Crawl. Such raw texts lack explicit representation of knowledge such as linguistic knowledge and world knowledge. In addition, most large-scale models are trained in an auto-regressive way, but [6] shows that such models demonstrate poorer performance with traditional fine-tuning when adapting to downstream language understanding tasks.

In this work, to solve the problem caused by a single auto-regressive framework and to explore the performance of knowledge enhanced pre-trained models with large-scale parameters, we propose a unified framework called ERNIE 3.0 to train large-scale knowledge enhanced models on a 4TB corpus consisting of plain texts and a large-scale knowledge graph by fusing the auto-regressive network and the auto-encoding network. The proposed ERNIE 3.0 can handle both natural language understanding tasks and natural language generation tasks through zero-shot learning, few-shot learning or fine-tuning. Furthermore, the proposed framework supports the introduction of various customized tasks at any time. These tasks share the same encoding networks and are trained through multi-task learning. This method makes the encoding of lexical, syntactic and semantic information across different tasks possible. Moreover, when given a new task, our framework could incrementally train the distributed representations based on the previous training parameters, with no need to train them from scratch.

In summary, our contributions are as follows:

- We propose a unified framework ERNIE 3.0, which combines auto-regressive network and auto-encoding network so that the trained model can handle both natural language understanding and generation tasks through zero-shot learning, few-shot learning or fine-tuning.
- We pre-train large-scale knowledge enhanced models with 10 billion parameters and evaluate them with a series of experiments on both natural language understanding and natural language generation tasks. Experimental results show that ERNIE 3.0 consistently outperforms the state-of-the art models on 54 benchmarks by a large margin and achieves the first place on the SuperGLUE [3] benchmark.

2 Related Work

2.1 Large-scale Pre-trained Models

Since BERT [6] is proposed as a powerful language model for natural language understanding, pre-trained language models have attracted more and more attention and become the new paradigm for natural language processing. One of the research trends is increasing model size, which leads to lower perplexity and better performance [14]. As a result, many large-scale pre-trained models have been proposed in the past two years. T5 model [1] is proposed to push the performance for both natural language understanding and natural language generation tasks with 11 billion parameters. The T5 model converts all text-based language tasks into a text-to-text format by a unified framework and fully explores the effectiveness of pre-training objectives, architectures, unlabeled datasets, transfer approaches, and other factors. After the T5 model, GPT-3 [2], which includes 175 billion parameters, is proposed to achieve an amazing performance on a wide range of tasks under the few-shot and zero-shot settings. Specifically, GPT-3 is an auto-regressive language model, 10x more than its predecessor, GPT-2, proposed by [15]. However, GPT-3 shows a lack of common sense, exists biases and privacy issues in the tests [16]. [13] have proposed a 1 trillion parameters model named Switch Transformer with simplifying MoE [17, 18] routing algorithm to improve model with less communication and computational costs, and [13] also proposed a large scale distributed training solution to tackle the problem of training complexity, communication costs, and training instability.

Besides the models mentioned above, more non-English large models have been proposed recently. [19] released a 2.6 billion parameters Chinese Pre-trained Language Model (CPM) with generative pre-training on large-scale Chinese training data and the model structure was inspired by [2]. [20] have released a 11 billion parameters model CPM-2. To accelerate the pre-training based on existing PLMs instead of training models from scratch, the knowledge inheritance techniques have been introduced and during the fine-tuning stage, prompt tuning is involved to better exploit the knowledge within the pre-trained model. [21] have proposed a cross-modal pre-training method called M6(Multi-Modality to Multi-Modality Multitask Mega-Transformer) including 100 billion parameters for unified pre-training on multiple modalities data. [22] proposed a 200 billion parameters auto regressive language model named PangGu-α which is trained on a cluster of 2048 Ascend 910 AI processors with distributed training techniques including data parallelism, op-level model parallelism, pipeline model parallelism, optimizer model parallelism and

re-materialization. Except for those Chinese large-scale models, a Korean 204 billion parameters language model named HyperCLOVA [23] has been proposed, and its volume of machine-learned data in Korean was 6,500 times larger than GPT-3's. From what has been discussed above, observations now suggest that large-scale pre-trained models have attracted more and more attention from industry and academia.

2.2 Knowledge Enhanced Models

Pre-trained language models capture syntactical and semantic knowledge from large-scale corpus, but lack world knowledge. Recently, several works have attempted to incorporate world knowledge in pre-trained language models. The typical form of world knowledge is a knowledge graph. Many works ([24, 25, 26]) integrate entity and relation embedding from knowledge graph in pre-trained language models. WKLM [27] replaced entity mentions in the original documents with names of other entities of the same type and train the models to distinguish the correct entity mention from randomly chosen ones. KEPLER [28] optimized the models with knowledge embedding and mask language model objectives to align the world knowledge and language representation into the same semantic space. CoLAKE [29] integrated the language context and the knowledge context in a word-knowledge graph and jointly learned contextualized representation for language and knowledge with the extended mask language model objective. Another existing form of world knowledge is the extra annotation of large-scale data. ERNIE 1.0 [7] introduced phrase masking and named entity masking and predicts the whole masked phrases and named entities to help the model learn the dependency information in both local contexts and global contexts. CALM [30] taught models to detect and revised a corrupted sentence with the incorrect ordering of concepts and to distinguish truth sentences from less plausible ones via two kinds of self-supervised pre-training tasks. K-Adapter[31] utilized adapters trained on different knowledge sources with extra annotations to distinguish where the knowledge comes from.
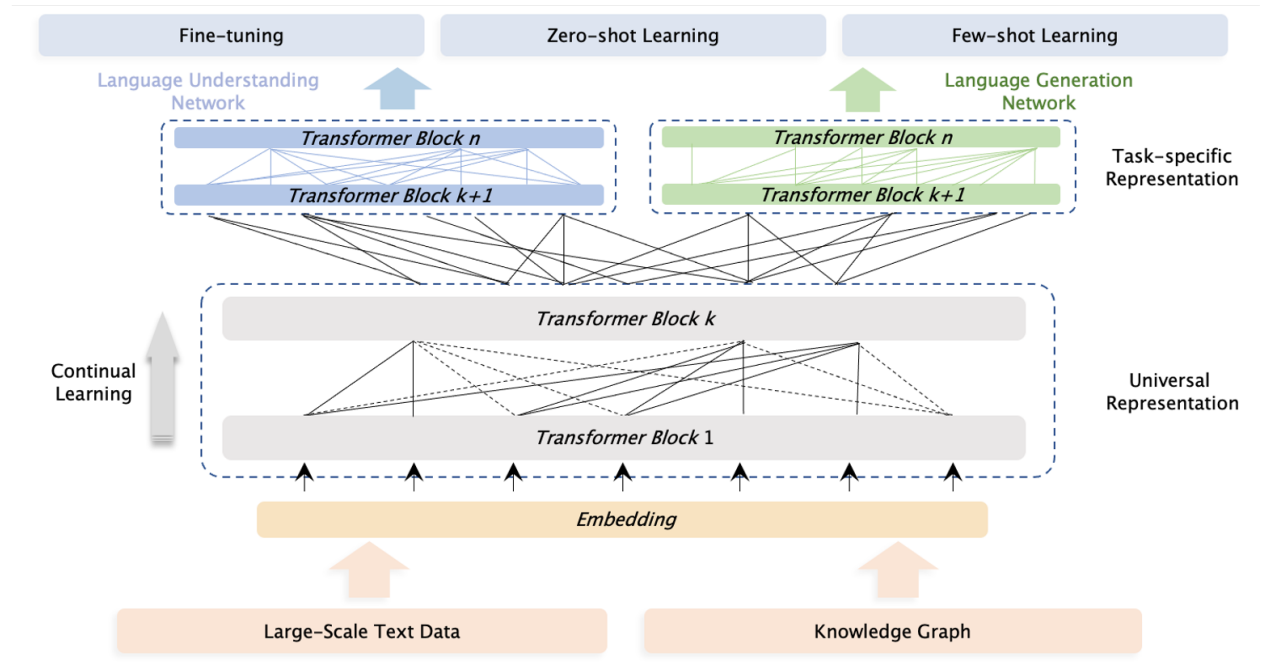
3 ERNIE 3.0

Figure 1: The framework of ERNIE 3.0.

A significant improvement has been achieved on various natural language processing tasks for knowledge enhanced pre-trained models with the base or large model size, such as ERNIE, ERNIE 2.0 and SpanBERT [32], in which the base/large model size represent 12/24 layers Transformer respectively. In order to explore the effectiveness of knowledge enhanced large-scale pre-trained model, we propose the ERNIE 3.0 framework to pre-train model on massive unsupervised corpus including plain texts and knowledge graph. Furthermore, we employ various types of pre-training tasks to enable the model to learn the different levels of knowledge consisting of valuable lexical, syntactic

and semantic information more effectively, in which the pre-training tasks spread three task paradigms, that is natural language understanding, natural language generation and knowledge extraction. Therefore, ERNIE 3.0 innovatively designs a **Continual Multi-Paradigms Unified Pre-training Framework** to enable the collaborative pre-training among multi-task paradigms. The explicit introduction of ERNIE 3.0 will be explained in the following sections.

3.1 Overview of ERNIE 3.0 Framework

The Framework of the ERNIE 3.0 is shown in Figure 1, which can be widely used for pre-training, fine-tuning and zero/few-shot learning. Unlike the prevalent unified pre-training strategy of employing a shared Transformer network for different well-designed cloze tasks and utilizing specific self-attention masks to control what context the prediction conditions on, ERNIE 3.0 designs a new *Continual Multi-Paradigms Unified Pre-training Framework*. We believed that the different task paradigms of natural language processing depend on identical underlying abstract features consistently, such as lexical information and syntactic information, but the requirements of top-level concrete features are incompatible, in which the natural language understanding tasks have the disposition to learn the semantic coherence while natural language generation tasks expect further contextual information. Therefore, inspired by the classical model architecture of multi-task learning, in which the lower layers are shared across all tasks while the top layers are task-specific, we proposed the ERNIE 3.0 to enable the different task paradigms to share the underlying abstract features learned in a shared network and utilizing the task-specific top-level concrete features learned in their own task-specific network respectively. Furthermore, in order to help the model efficiently learn the lexical, syntactic and semantic representations, ERNIE 3.0 exploits the continual multi-task learning framework introduced in ERNIE 2.0 [33]. As for the application of different kinds of downstream tasks, we will first initialize the ERNIE 3.0 with the combination of parameters of a pre-trained shared network and corresponding task-specific networks for different task paradigms, and then execute the corresponding follow-up procedure using data from specific tasks.

We refer to the backbone shared network and task-specific networks as the **Universal Representation Module** and **Task-specific Representation Module**s in ERNIE 3.0. Specifically, the universal representation network plays the role of universal semantic features extractor (for example, it can be a multi-layer Transformer), in which the parameters are shared across all kinds of task paradigms, including natural language understanding, natural language generation and so on. And the task-specific representation networks undertake the function of extracting the task-specific semantic features, in which the parameters are learned by task-specific objectives. ERNIE 3.0 not only enables the model to distinguish the task-specific semantic information among different task paradigms, but also mitigates the dilemma that large-scale pre-trained models are difficult to implement with limited time and hardware resources, in which ERNIE 3.0 permits the models to only update the parameters of a task-specific representation network during the fine-tuning phase. Specifically, ERNIE 3.0 employs the collaborative architecture of a Universal Representation Module and two Task-specific Representation Modules, namely natural language understanding (NLU) specific representation module and natural language generation (NLG) specific representation module.

3.1.1 Universal Representation Module

ERNIE 3.0 uses a multi-layer Transformer-XL [34] as the backbone network like other pre-trained models such as XLNet [35], Segatron [36] and ERNIE-Doc [37], in which Transformer-XL is similar to Transformer but introduces an auxiliary recurrence memory module to help modelling longer texts. We refer to the backbone as Universal Representation Module and it is shared across all the task paradigms. Proverbially, the Transformer can capture the contextual information for each token in the sequence via self-attention and generate a sequence of contextual embedding. It is evident that the larger the scale of Transformer model, the stronger its capacity to capture and store up various semantic information with different levels. Therefore, ERNIE 3.0 sets the universal representation module with a larger size to enable the model to effectively capture universal lexical and syntactic information from training data by learning various pre-training tasks of different paradigms. And what needs special attention is that the memory module is only valid for natural language generation tasks while controlling the attention mask matrices.

3.1.2 Task-specific Representation Module

Similar to the basic shared representation module, the task-specific representation module is also a multi-layer Transformer-XL, which is used to capture the top-level semantic representations for different task paradigms. ERNIE 3.0 sets the task-specific representation module to a manageable size, that is a base model size, instead of the multi-layer perceptron or shallow Transformer commonly used in multi-task learning, which will produce three obvious benefits, the first is that the base network has a stronger ability to capture semantic information than multi-layer perceptron and shallow Transformer; the second is that the task-specific networks with base model size enable ERNIE 3.0 to distinguish the top-level semantic information among different task paradigms without significantly increasing the parameters of a large-scale model; finally, the smaller model size of a task-specific network than a shared network

would lead to realizable practical applications for large scale pre-trained model when only fine-tuning on the task-specific representation module. ERNIE 3.0 constructs two task-specific representation modules, that is NLU-specific representation module and NLG-specific representation module, in which the former is a bi-directional modeling network while the latter is a uni-directional modeling network.

3.2 Pre-training Tasks

We construct several tasks for various task paradigms to capture different aspects of information in the training corpora and make the capacity of understanding, generation and reasoning available to pre-trained model.

3.2.1 Word-aware Pre-training Tasks

Knowledge Masked Language Modeling   ERNIE 1.0 [7] proposed an effective strategy to enhance representation through knowledge integration, namely Knowledge Integrated Masked Language Modeling task. It introduced phrase masking and named entity masking that predict the whole masked phrases and named entities to help the model learn the dependency information in both local contexts and global contexts.

Document Language Modeling   Generative pre-training models usually utilize traditional language model (such as GPT [5], GPT-2 [15]) or sequence-to-sequence language model (such as BART [38], T5 [1], ERNIE-GEN [39]) as the pre-training task, the latter trains on the network with an auxiliary decoder structure. ERNIE 3.0 opt for traditional language model as the pre-training task to abate the network complexity and heighten the effectiveness of unified pre-training. In addition, to enable the NLG network of ERNIE 3.0 to model longer text, we introduce the Enhanced Recurrence Memory Mechanism proposed in ERNIE-Doc [37], which can model a larger effective context length than traditional recurrence Transformer by changing the shifting-one-layer-downwards recurrence to the same-layer recurrence.

3.2.2 Structure-aware Pre-training Tasks

Sentence Reordering   Sentence reordering task, which is introduced in ERNIE 2.0 [29], aims to train the model to learn the relationship between sentences by reorganizing permuted segments. At length, a given paragraph is randomly split into 1 to m segments during pre-training and all of the combinations are shuffled by a random permuted order. Then, the pre-trained model is asked to reorganize these permuted segments, modeled as a k-class classification problem where $k = \sum_{n=1}^{m} n!$.

Sentence Distance   Sentence distance task, an extension of traditional next sentence prediction (NSP) task, is widely used in various pre-trained models to enhance their ability to learn the sentence-level information, which can be modeled as a 3-class classification problem. The three categories represent that the two sentences are adjacent, nonadjacent but in the same document and from two different documents respectively.
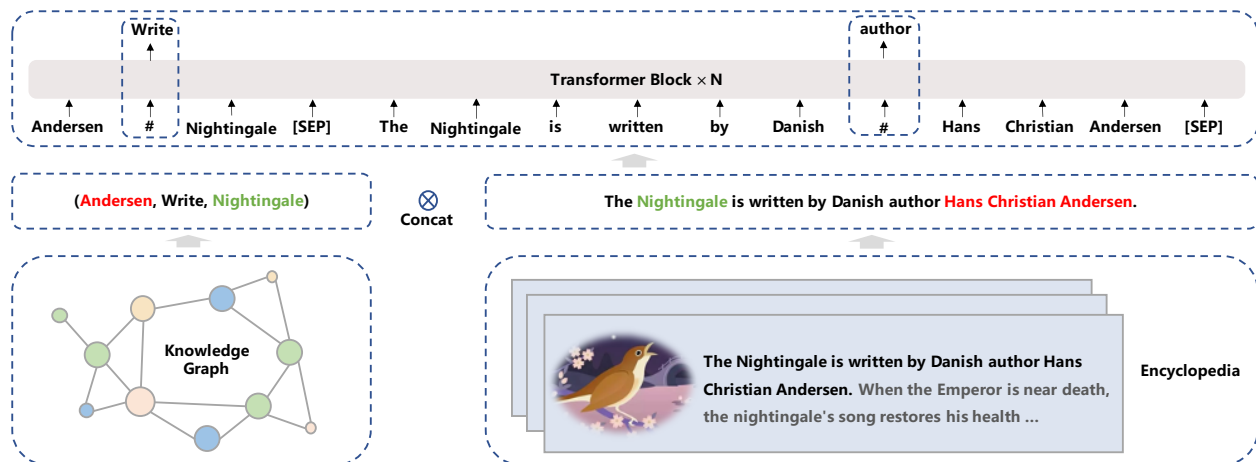

Figure 2: Universal Knowledge-Text Prediction.

5

Corpus	ERNIE 2.0	Search	Web	QA-long	QA-short	Novel	Poetry&Couplet	Medical	Law	Fin	KG
# of tokens	17.8B	42.4B	314.7B	33.8B	0.1B	96.4B	46.5M	17.8B	16.2B	0.6B	0.7B
multiplier	20	7	1	3	40	1	20	1	1	10	10
# tokens of context length in each percentile using ERNIE-3.0 wordpiece tokenizer											
50%	135	75	793	184	15	2063	30	314	1162	843	16
95%	1257	827	2757	1168	22	3652	88	983	4587	1572	44

Table 1: Statistics of Pre-training Datasets.

3.2.3 Knowledge-aware Pre-training Tasks

Universal Knowledge-Text Prediction To incorporate knowledge into one pre-trained language model, we introduce universal knowledge-text prediction (UKTP) task, which is an extension of knowledge masked language modeling. While knowledge masked language modeling only requires unstructured texts, universal knowledge-text prediction task requires both unstructured texts and knowledge graphs. The universal knowledge-text prediction task is illustrated in Figure 2. Given a pair of triple from knowledge graph and the corresponding sentence from encyclopedia, we randomly mask relation in triple or words in a sentence. To predict the relation in the triple, the model needs to detect mentions of head entity and tail entity and determine semantic relationship that holds between them in the corresponding sentence. The essence of this process is similar to the distant supervision algorithm [40] in relation extraction tasks. The distant supervision algorithm assume that if two entities participate in a relation, any sentence that contain those two entities might express that relation. Meanwhile, to predict words in the corresponding sentence, the model not only considers the dependency information in the sentence, but also logical relationship in the triple. Specifically, the procedure of obtaining pairs of a triple and this corresponding sentence is as follows: given a document from encyclopedia, we first find the candidate triples in the knowledge graph whose mentions of head entity or tail entity is title of the document, and then select triples from candidate triples whose mentions of head entity and tail entity are mentioned in the same sentence in the document.

ERNIE 3.0 trains the NLU network through knowledge masked language modeling to improve the capacity of capturing the lexical information, trains the sentence reordering task and the sentence distance discerning task to strengthen the ability of capturing the syntactic information, and finally optimizes the model with the universal knowledge-text prediction task to improve knowledge memorization and reasoning. Meanwhile, ERNIE 3.0 trains the NLG network with the document language modeling task to enable various generation styles.

3.3 Pre-training Process

3.3.1 Pre-training Algorithm

Progressive training was originally proposed to improve stability, which starts from an efficient and small model and gradually increase the capacity [41]. Recent study leverages this paradigm to accelerate model training. As large-scale pre-training keeps advancing the state-of-the-art([6], [5]), their overwhelming computational consumption becomes the major burden towards further developing more powerful models([15]). Preliminary application of progressive training has been made on Transformer pre-training. BERT([6]) designs a two-stage training with a reduced sequence length for the first 90% of updates. [15] also gradually increase the batch size linearly from a small value to the full value. [42] also notice that changing the regularization factors (e.g. [43], [44]) stage-wise with respect to the input size can speed up training networks. To further improve convergence speed of the training process, we propose to adjust the training regularization factors in a more comprehensive and smooth way by progressively and simultaneously increasing the training factors including the input sequence length, the batch size, the learning rate and the dropout rate. In fact, it is common that Transformer models adopts the learning rate warm-up strategy to increase training stability and our improved progressive learning strategy is compatible to the existing strategy.

3.3.2 Pre-training Data

To ensure the success of the pre-training of ERNIE 3.0, we construct a large-scale, wide-variety and high-quality Chinese text corpora amounting to 4TB storage size in 11 different categories. To our best knowledge, this is currently the largest Chinese pre-training corpora compared with CLUECorpus2020 [45] (100GB), Chinese multi-modal pre-training data [21] (300GB), WuDaoCorpus2.0 used by CPM-2 [20] (2.3TB Chinese data and 300GB English data) and PanGu Corpus [22] (1.1TB).

In detail, we build the corpus for ERNIE 3.0 based on that from ERNIE 2.0 (including baike, wikipedia, feed and etc), Baidu Search (including Baijiahao, Zhidao, Tieba, Experience), Web text, QA-long, QA-short, Poetry [2] &Couplet [3], Domain-specific data from medical, law and financial area and Baidu knowledge graph with more than 50 million facts. To improve the data quality, we adopt the following pre-processing strategies:

- Deduplication is conducted on different granularities including character level, paragraph level and document level. On the character level, we replace consecutive identical characters (i.e., spaces, tabs, exclamation mark, question mark and etc) with one single character. One the paragraph level, we replace two identical consecutive paragraphs consisting of N sentences with one single paragraph where $0 < N < 100$. The two aforementioned deduplication strategies are critical for ERNIE 3.0 to generate non-repeating contents. At last, we adopted Message Digest Algorithm5 (MD5) to filter duplicate documents by comparing the sum of the MD5 of top-3 longest sentences from each document.

- Sentences with less than 10 words are filtered since they may be problematic or incomplete ones which contains limited semantic information for model pre-training.

- We further conduct sentence segmentation using regular expressions and word segmentation based on Baidu's word segmentation tool. This helps ERNIE 3.0 to learn better sentence boundary and named entity knowledge during pre-training.

Then, each dataset is multiplied by a user-defined multiplier number to increase the data diversity after truncating the data for NLU-network pre-training.

3.3.3 Pre-training Settings

Both the universal representation module and the task-specific representation modules of ERNIE 3.0 uses the Transformer-XL[34] structure as the backbone. For the universal representation module, we adopt a structure with 48 layers, 4096 hidden units and 64 heads. For the task-specific representation modules, we adopt a structure with 12 layers, 768 hidden units and 12 heads. The total parameter of universal representation module and task-specific representation modules is 10 billion. The activation function used is GeLU[46]. The maximum sequence length of context and the memory length of language generation is set to 512 and 128, respectively. The total batch size of all pre-training tasks is set to 6144. We use Adam[47] with learning rate of 1e-4, $\beta_1 = 0.9$, $\beta_2 = 0.999$, L2 weight decay of 0.01, learning rate warmup over the first 10,000 steps and linear decay of the learning rate. In the first 10,000 steps, we also use the progressive learning to speedup convergence in the initial stage of pre-training. The model is trained for a total of 375 billion tokens with 384 NVDIA v100 GPU cards and is implemented on PaddlePaddle framework. By virtue of parameter sharding used in [48, 49], we manage to reduce the memory usage of our model and address the problem of the total parameter of model exceeding the memory of a single GPU card.

4 Experiments

We compare the performance of ERNIE 3.0 with the state-of-the-art [4] pre-training models through fine-tuning on both natural language understanding tasks (in Sec. 4.2.1) and natural language generation tasks (in Sec. 4.2.2), and zero-shot learning (in Sec. 4.3)[5].

4.1 Evaluation Tasks

We executed extensive experiments on 54 NLP tasks to evaluate the fine-tuning and zero-shot learning performances of the models.

4.1.1 Natural Language Understanding Tasks

45 datasets belonging to 14 kinds of natural language understanding tasks are used in our experiments, as follows:

---

[2]https://www.luge.ai/text-generation/chinese-poetry.html#_1-chinese-poetry

[3]https://github.com/v-zich/couplet-clean-dataset

[4]the previous state-of-the-art results are all from the public single model that we can find.

[5]The previous SoTA results of ERNIE 2.0 and RoBERTa-wwm-ext on corresponding datasets are reproduced by ourselves, except for the datasets that already have released pre-trained results.

- **Sentiment Analysis**: NLPCC2014-SC [6], SE-ABSA16_PHNS [7], SE-ABSA16_CAME, BDCI2019 [8].
- **Opinion extraction**: COTE-BD [50], COTE-DP [50], COTE-MFW [50].
- **Natural Language Inference**: XNLI [51], OCNLI [45], CMNLI [45].
- **Winograd Schema Challenge** CLUEWSC2020 [45].
- **Relation Extraction**: FinRE [52], SanWen [53].
- **Event Extraction**: CCKS2020 [9].
- **Semantic Similarity**: AFQMC [45], LCQMC [54], CSL [45], PAWS-X [55], BQ Corpus [56].
- **Chinese News Classification**: TNEWS [10], IFLYTEK [57], THUCNEWS [11], CNSE [58], CNSS [58].
- **Closed-Book Question Answering**: NLPCC-DBQA [12], CHIP2019, cMedQA [59], cMedQA2 [60], CK-BQA [13], WebQA [61].
- **Named Entity Recognition**: CLUENER [45], Weibo [62], OntoNotes [63], CCKS2019 [14].
- **Machine Reading Comprehension**: CMRC 2018 [64], CMRC2019 [65], DRCD [66], DuReader [67], Dureader$_{robust}$ [68], Dureader$_{checklist}$, Dureader$_{yesno}$ [15], C^3 [69], CHID [70].
- **Legal Documents Analysis**: CAIL2018-Task1 [71], CAIL2018-Task2 [71].
- **Cant Understanding**: DogWhistle Insider, DogWhistle Outsider[72].
- **Document Retrieval**: Sogou-log [73].

4.1.2 Natural Language Generation Tasks

9 datasets belonging to 7 kinds of natural language generation tasks are used in our experiments, as follows:

- **Text Summarization**: LCSTS [10]
- **Question Generation**:KBQG [16], DuReader-QG [67], DuReader$_{robust}$-QG [68].
- **Closed-Book Question Answering**: MATINF-QA [74].
- **Math**: Math23K [75].
- **Advertisement Generation**: AdGen [76].
- **Translation**: WMT20-enzh [77].
- **Dialogue Generation**: KdConv [78].

4.2 Experiments on Fine-tuning Tasks

4.2.1 Fine-tuning on Natural Language Understanding Tasks

The results of natural language understanding tasks are reported in Table 2.

Sentiment Analysis. Sentiment Analysis is a classification task aiming to determine whether a sentence is positive, negative, or neutral. We consider 4 datasets from different domains, including shopping (NLPCC2014-SC), electronics (SE-ABSA16_PHNS, SE-ABSA16_CAM), and financial (BDCI2019). ERNIE 3.0 achieves a substantial improvement on all four datasets.

Opinion Extraction. Similar to the sentiment analysis task, opinion extraction requires the model to mine the opinion of a sentence. We use 3 sub-datasets from Chinese Customer Review (COTE). Experiment results show that ERNIE 3.0 also outperforms the current SoTA system by a great margin.

---

[6] http://tcci.ccf.org.cn/conference/2014/pages/page04_dg.html
[7] http://alt.qcri.org/semeval2016/task5/
[8] https://www.datafountain.cn/competitions/350
[9] http://sigkg.cn/ccks2020/?page_id=69
[10] https://github.com/aceimnorstuvwxz/toutiao-text-classfication-dataset
[11] http://thuctc.thunlp.org/
[12] http://tcci.ccf.org.cn/conference/2016/dldoc/evagline2.pdf
[13] https://github.com/pkumod/CKBQA
[14] https://www.biendata.xyz/competition/ccks_2019_1/
[15] https://aistudio.baidu.com/aistudio/competition/detail/49/?isFromLUGE=TRUE
[16] http://tcci.ccf.org.cn/conference/2017/dldoc/taskgline05.pdf

ID	Task	Dataset	Metric		Previous SoTA Model	ERNIE 3.0
1	Sentiment Analysis	NLPCC2014-SC	Acc.	Test	83.53 (SKEP)	**86.00**
		SE-ABSA16_PHNS	Acc.	Test	82.91 (SKEP)	**93.95**
		SE-ABSA16_CAME	Acc.	Test	90.06 (SKEP)	**96.05**
		BDCI2019	Acc.	Dev	-	96.83
				Test	96.26 (ERNIE 2.0)	**97.70**
2	Opinion Extraction	COTE-BD	F1	Test	84.50 (SKEP)	**90.23**
		COTE-DP	F1	Test	86.30 (SKEP)	**92.75**
		COTE-MFW	F1	Test	87.90 (SKEP)	**89.90**
3	Natural Language Inference	OCNLI	Acc.	Dev	78.80 (RoBERTa*)	**82.75**
		XNLI	Acc.	Dev	83.25 (Zen 2.0)	**84.42**
				Test	83.09 (Zen 2.0)	**83.77**
4	Winograd Schema Challenge	WSC2020	Acc.	Dev	69.70 (RoBERTa*)	**95.40**
5	Relation Extraction	FinRE	F1	Dev	63.33 (ERNIE 2.0)	**64.87**
				Test	60.60 (ERNIE 2.0)	**62.88**
		SanWen	F1	Dev	79.92 (ERNIE 2.0)	**81.32**
				Test	77.97 (ERNIE 2.0)	**82.59**
6	Event Extraction	CCKS2020	F1	Dev	60.64 (ERNIE 2.0)	**61.70**
				Test	61.34 (ERNIE 2.0)	**64.33**
7	Semantic Similarity	AFQMC	Acc.	Dev	74.92 (RoBERTa*)	**77.02**
		LCQMC	Acc.	Dev	-	90.29
				Test	89.16 (CPM-2)	**90.38**
		CSL	Acc.	Dev	82.17 (RoBERTa*)	**84.50**
		PAWS-X	Acc.	Dev	86.25 (ERNIE 2.0)	**87.00**
				Test	86.35 (ERNIE 2.0)	**87.10**
		BQ Corpus	Acc.	Dev	87.11 (ZEN 2.0)	**87.41**
				Test	85.99 (ZEN 2.0)	**86.10**
8	Chinese News Classification	TNEWS	Acc.	Dev	58.32 (RoBERTa*)	**69.94**
		IFLYTEK	Acc.	Dev	62.75 (RoBERTa*)	**63.45**
		THUNCEWS	Acc.	Dev	97.7 (RoBERTa*)	**98.33**
				Test	97.6 (RoBERTa*)	**98.66**
		CNSE	Acc.	Dev	85.64 (RoBERTa*)	**88.94**
				Test	85.57 (RoBERTa*)	**88.92**
		CNSS	Acc.	Dev	93.06 (ERNIE 2.0)	**93.84**
				Test	92.73 (ERNIE 2.0)	**93.76**
9	Closed-Book Question Answering	NLPCC-DBQA	MRR/F1	Dev	96.04/85.69 (Zen 2.0)	**96.71/87.57**
				Test	96.11/86.47 (Zen 2.0)	**96.50/88.49**
		CHIP2019	Acc.	Test	89.22 (ERNIE 2.0)	**89.90**
		cMedQA	Acc.	Dev	78.6 (BERT_BiGRU*)	**84.60**
				Test	78.2 (BERT_BiGRU*)	**82.65**
		cMedQA2	Acc.	Dev	81.3 (BERT_BiGRU*)	**83.48**
				Test	82.2 (BERT_BiGRU*)	**83.68**
10	Named Entity Recognition	CLUENER	F1	Dev	80.42 (RoBERTa*)	**81.23**
		Weibo	F1	Dev	-	70.06
				Test	67.60 (Glyce+BERT)	**69.23**
		OntoNotes	F1	Dev	-	79.59
				Test	81.63 (Glyce+BERT)	**82.64**
		CCKS2019	F1	Test	81.58 (ERNIE 2.0)	**82.70**
11	Cant Understanding	DogWhistle Insider	Acc.	Dev	75.4 (ALBERT)	**79.06**
				Test	76.1 (ALBERT)	**79.22**
		DogWhistle Outsider	Acc.	Dev	34.6 (ALBERT)	**38.68**
				Test	34.6 (ALBERT)	**38.22**

ID	Task	Dataset	Metric		Previous SoTA Model	ERNIE 3.0
12	Machine Reading Comprehension	CMRC2018	EM/F1	Dev	74.3/90.5 (ERNIE-Gram)	**75.30/92.29**
		CRMC2019	QAC/PAC	Dev	82.6/23.3 (RoBERTa*)	**92.53/57.33**
		DRCD	EM/F1	Dev	90.8/95.3 (MacBERT)	**91.54/96.45**
				Test	90.9/95.3 (MacBERT)	**91.41/95.84**
		DuReader	EM/F1	Dev	64.2/77.3 (ERNIE 2.0)	**67.69/79.66**
		DuReader$_{robust}$	EM/F1	Dev	75.23/86.77 (ERNIE 2.0)	**77.27/88.54**
				Test	51.20/67.96 (ERNIE 2.0)	**60.87/75.63**
		DuReader$_{checklist}$	EM/F1	Dev	55.66/64.12 (ERNIE 2.0)	**61.33/70.59**
				Test	59.11/48.79 (ERNIE 2.0)	**64.87/53.82**
		DuReader$_{yesno}$	Acc.	Dev	88.69 (ERNIE 2.0)	**89.95**
				Test	88.82 (ERNIE 2.0)	**89.64**
		C3	Acc.	Dev	-	87.63
				Test	86.1 (CPM-2)	**86.69**
		CHID	Acc.	Dev	85.81 (RoBERTa*)	**91.67**
13	Legal Document Analysis	CAIL2018 Task1	F1-macro/F1-micro	Dev	83.85/91.50 (ERNIE 2.0)	**88.64/93.11**
				Test	80.40/89.94 (ERNIE 2.0)	**86.83/91.82**
		CAIL2018 Task2	F1-macro/F1-micro	Dev	78.58/89.46 (ERNIE 2.0)	**82.62/90.93**
				Test	75.35/86.97 (ERNIE 2.0)	**81.10/88.52**
14	Document Retrieval	Sogou-log	MRR/NDCG@1	Test	36.3/35.5 (CPM-2)	**38.20/37.24**

Table 2: Results on Natural Language Understanding Tasks. We compare ERNIE 3.0 with 10 previous SoTA baselines including CPM-2[20], ERNIE 2.0[33], ERNIE-Gram[79], SKEP[80], RoBERTa-wwm-ext-large[81] (marked as RoBERTa*), ALBERT[82], MacBERT[83], Zen 2.0[84], Glyce[85] and crossed BERT siamese BiGRU[86] (marked as BERT_BiGRU*).

Natural Language Inference. Natural Language Inference is the task to determine whether a given premise semantically entails another hypothesis. We use OCNLI and XNLI datasets. The results indicate that ERNIE 3.0 has achieved 3.9 and 0.7 accuracy improvement on two datasets, respectively. The improvement on the XNLI dataset is quite limited, and this may be due to the poor quality of the dataset since the XNLI dataset is translated from English.

Winograd Schemas Challenge. WSC2020 is an anaphora resolution task where the model is asked to decide whether a pronoun and a noun in a sentence co-refer, ERNIE 3.0 achieves a significant improvement of 25.7 points.

Relation Extraction. The task of relation extraction is to identify the relationship between different entities like persons and organizations. We consider FinRE and SanWen – two relation extraction datasets for financial news and Chinese literature respectively. ERNIE 3.0 outperforms the previous SoTA model by 2.46 points on average.

Event Extraction. Similar to relation extraction, the event extraction task aims to identify the event entities and classify them into different categories. We choose CCKS2020 – a text-level event subject extraction dataset of financial field. ERNIE 3.0 has 3 points of improvement on the test set.

Semantic Similarity. Semantic Similarity is a classic NLP task that determines the similarity between various terms such as words, sentences, documents. In this work, we focus on sentence-level similarity tasks. We test ERNIE 3.0 on several datasets in varied fields including AFQMC, LCQMC, CSL, PAWS-X, and BQ. Experiment results show that ERNIE 3.0 outperforms the baseline models by a remarkable margin. Especially, under comparable number of parameters, ERNIE 3.0 surpasses CPM-2 with 1.2 points on LCQMC dataset.

Chinese News Classification. We also evaluate ERNIE 3.0 on Chinese news classification. We consider 6 datasets including news title (TNEWS), app descriptions (IFLYTEK), and news stories (THUCNEWS, CNSE, CNSS). Under different types of classification tasks, ERNIE 3.0 can consistently achieve better accuracy with 2.8 points improvement on average.

Closed-Book Question Answering. Closed-Book Question Answering aims to directly answer the questions without any additional references or knowledge. We select a general QA dataset NLPCC-DBQA and three medical field datasets – CHIP2019, cMedQA, and cMedQA2 to test the ability of ERNIE 3.0. Experiment results show that ERNIE 3.0 performs better on all QA tasks, we believe knowledge enhanced pre-training methods do bring benefits to the closed-book QA task.

Cant Understanding. Cant, also known as doublespeak, is an advanced language usage for humans. However, it is rather difficult for machines to understand this type of language. We test the cant understanding ability of ERNIE 3.0 on DogWhistle – a dataset based on *Decrypto* game. The model is required to select the right answer with the guidance of the corresponding cant. ERNIE 3.0 gets the best result and shows its potential for understanding much more difficult languages.

Task	Dataset	Metric	RoBERTa-Large	ERNIE 2.0-Large	ProphetNet-zh	mT5	CPM-2	ERNIE 3.0
Text Summarization	LCSTS	ROUGE-L	40.98	41.38	37.08	34.8	35.88	**48.46**
Question Generation	KBQG	BLEU-4	-	57.40	-	-	-	**64.70**
	DuReader-QG	BLEU-4	32.29	34.15	-	-	-	**48.36**
	DuReader$_{robust}$-QG	BLEU-4	37.10	39.30	-	-	-	**41.70**
Closed-Book Question Answering	MATINF-QA	ROUGE-L	-	-	15.47	-	-	**17.33**
Math	Math23K	Acc.	-	-	-	61.60	69.37	**75.00**
Advertisement Generation	AdGen	BLEU-4	-	-	-	9.82	10.60	**30.16**
Translation	WMT20-enzh	BLEU	-	-	-	23.98	26.21	**26.80**
Dialogue Generation	KdConv	BLEU-4	15.75	13.94	-	-	-	**23.85**

Table 3: Results on Natural Language Generation Tasks. We reported the results on the test set.

Named Entity Recognition. Named Entity Recognition is a classical NLP task of extracting and classifying entities in text. We select widely used OntoNotes, CLUENER, Weibo, and a domain-specific dataset CCKS2019. From the results, ERNIE 3.0 performs better than the baseline models across all datasets.

Machine Reading Comprehension. We comprehensively evaluate the ability of ERNIE 3.0 on machine reading comprehension in different aspects, including span-predict reading comprehension (CMRC2018, DuReader, DRCD, DuReader$_{checklist}$), multiple-choice reading comprehension (C3, DuReader$_{yesno}$), cloze and completion (CHID, CMRC2019), and robustness test (Dureader$_{robust}$). With the help of knowledge enhanced pre-training, ERNIE 3.0 surpasses the baseline models with significant enhancements on all types of tasks. To be more specific, ERNIE 3.0 achieve at least 1.0 points of EM improvement on 5 span-predict tasks and 0.89 accuracy improvement on multiple-choice tasks on average. Also, under comparable number of parameters, ERNIE 3.0 outperforms CPM-2 with 0.6 points on C3 dataset. For the robustness test, ERNIE 3.0 also performs best on the test set with over-sensitivity and over-stability samples.

Legal Documents Analysis. Next, we test the ability of ERNIE 3.0 on document analysis, we choose two domain-specific tasks of law. These two datasets from CAIL2018 are both multi-label document classification tasks. ERNIE 3.0 outperforms ERNIE 2.0 with remarkable increment.

Document Retrieval. Document retrieval aims to match documents given queries. We evaluate the retrieval ability of ERNIE 3.0 on Sogou-Log. Following previous work [20], we report NDCG@1 performance on the test-same test set and MRR performance on the test-raw test set and ERNIE 3.0 outperforms CPM-2.

4.2.2 Fine-tuning on Natural Language Generation Tasks

The results of natural language generation tasks are reported in Table 3.

Text Summarization. We consider a Large Scale Chinese Short Text Summarization (LCSTS) dataset which requires a model to understand the text and refine the key information to generate coherent, informative summaries. LCSTS is a classic Chinese text summarization dataset which consists of 2 million real Chinese short texts with short summaries from Sina Weibo. ERNIE 3.0 achieves 48.46% Rouge-L score which outperforms CPM-2 with comparable number of parameters (11B) and current SoTA ProphetNet-zh.

Question Generation. Question Generation is the reverse task of Machine Reading Comprehension (MRC) which requires the model to understand a document and generate a reasonable question based on a given short answer. We use a suite of three datasets including knowledge base question generation (KBQG), two MRC datasets named Dureader and Dureader$_{robust}$. ERNIE 3.0 performs best on these three datasets compared to the baselines.

Math. To test ERNIE 3.0's ability to perform simple arithmetic operations, we consider the Math23K dataset which contains 23,161 real math word problems for elementary school students with problem descriptions, structured equations and answers. ERNIE 3.0 is fine-tuned to generate the postfix expression of the structured equation given the problem description, then the final answer can be calculated using the Python *eval()* function (note that the '[' and ']' should be replaced with '(' and ')' respectively, also the '%' should be replaced with '*0.01' to avoid the failed solutions using Python *eval()* function). It shows that ERNIE 3.0 is a great math solver which achieves high accuracy 75% compared to CPM-2 69.37%.

Advertisement Generation. We consider AdGen which consists of 119K pairs of advertising text and clothing specification tables from a Chinese e-commerce platform. It requires the model to generate a long advertising text that covers all given attribute-value pairs for a piece of clothing. An attribute-value pair is joined with a colon, and several attribute-value pairs are concatenated sequentially using a '|' according to their segment number. Then we take

Task Paradigm	Task	Dataset	Metric	RoBERTa-Large	ERNIE 2.0-Large	ERNIE 3.0
NLU	Sentiment Analysis	NLPCC14-SC	Acc.	83.56	84.36	**86.00**
	Machine Reading Comprehension	DuReader$_{robust}$	EM/F1	51.10/67.18	51.20/67.96	**60.87/75.63**
	Semantic Similarity	LCQMC	Acc.	87.40	87.90	**90.38**
NLG	Question Generation	DuReader$_{robust}$-QG	BLEU-4	37.10	39.30	**41.70**
	Text Summarization	LCSTS	Rouge-L	40.98	41.38	**48.46**
	Dialogue Generation	KdConv	BLEU-4	15.75	13.94	**23.85**
Average				53.99	54.41	**59.77**

Table 4: Results on the LUGE benchmark. We reported the results on the test set.

Task Type	Dataset	Metric	CPM-1	PanGu-α-2.6B	PanGu-α-13B	ERNIE 3.0
Chinese News Classification	TNEWS	Acc.	65.44	60.95	60.26	**68.40**
	IFLYTEK	Acc.	68.91	74.26	73.80	**75.34**
Semantic Similarity	AFQMC	Acc.	66.34	59.29	65.76	**68.99**
	CSL	Acc.	52.30	50.50	49.30	**55.63**
Natural Language Inference	OCNLI	Acc.	44.20	42.61	41.53	**44.31**
	CMNLI	Acc.	49.10	47.56	49.29	**49.41**
Winograd Schema Challenge	WSC2020	Acc.	73.68	73.36	75.00	**78.38**
Cloze and completion	CHID	Acc.	68.62	68.73	70.64	**77.78**
	PD	Acc.	35.73	38.47	43.84	**66.07**
	CFT	Acc.	38.99	42.39	46.60	**49.30**
	CMRC2017	Acc.	24.60	37.83	38.90	**56.66**
	CMRC2019	Acc.	47.69	61.93	68.19	**75.00**
	WPLC	PPL	-	48.98	45.85	**17.03**
Machine Reading Comprehension	C3	Acc.	49.81	53.42	**54.47**	52.62
	CMRC2018	EM/F1	0.59/10.12	1.21/16.65	1.46/19.28	**7.61/25.61**
	DRCD	EM/F1	0.00/4.62	0.80/9.99	0.66/10.55	**10.58/26.29**
	DuReader	EM/F1	16.63	21.07	24.46	**29.79**
Closed-book Question Answering	WebQA	EM/F1	6.00/12.59	4.43/13.71	5.13/14.47	**22.53/38.95**
	CKBQA	Acc.	13.40	14.61	14.21	**20.64**

Table 5: Results on zero-shot learning tasks.

the structural attribute-value pairs string as input for ERNIE 3.0. It shows that ERNIE 3.0 is capable to generate a coherent and intriguing long advertising text by extracting information from a structural input with 19.56 percent point improvement w.r.t BLEU-4 compared to CPM-2.

Translation. For ERNIE 3.0, we mainly consider the pre-training on Chinese corpus. To test its multilingual ability, we expand our vocabulary to include extra 10K English subwords. On a classic multilingual dataset WMT20-enzh, we fine-tuned ERNIE 3.0 to translate English to Chinese. Compared to mT5-xxLarge and CPM-2, ERNIE 3.0 [17] is the best and presents superior multilingual ability.

Dialogue Generation. Next, we evaluate ERNIE 3.0 on Dialog Generation task. We consider a Chinese multi-domain knowledge-driven conversation dataset that contains 4.5K conversations from three domains (film, music, and travel). We train and test ERNIE 3.0 on the fused set of data from aforementioned three domains by only giving dialogue history to generate the current utterance. Knowledge triplets are excluded from inputs, so it's suitable to test a model's ability to model multi-turn conversations by leveraging inherent knowledge during pre-training. Compared to baselines, ERNIE 3.0 improves the performance a lot by 8.1 percent point, and we believe the knowledge graph enhanced pre-training attributes a lot.

---

[17] Due to the large size of the training dataset of WMT20-enzh, ERNIE 3.0 is not fully trained to convergence. We reported the BLEU score at 1.5 epoch checkpoint using SacreBLEU project [87].

4.2.3 LUGE benchmark

In order to further evaluate the capabilities of different models comprehensively and conveniently, we conduct experiments on the Language Understanding and Generation Evaluation Benchmarks(LUGE)) [18]. We use six representative tasks (see Tab. 4) from LUGE. ERNIE 3.0 delivers an average 5.36 percent improvement over leading pre-trained models such as ERNIE 2.0 and RoBERTa.

4.3 Experiments on Zero-shot Learning

We have demonstrated that ERNIE 3.0 is superior to previous SoTA methods on both NLU and NLG tasks following the pretraining-then-finetuning paradigm. In this section, we conduct various types of tasks with the zero-shot setting where a model is applied without any gradient updates or fine-tuning. ERNIE 3.0 achieves strong performance compared to recently proposed large-scale language models such as CPM-1 (2.6B), PanGu-α-2.6B and PanGu-α-13B on most downstream tasks. At last, we show that ERNIE 3.0 can generate more coherent, natural and accurate responses rated on our manually collected 450 cases across 13 different tasks.

4.3.1 Evaluation

The evaluation methods can be classified into two categories, namely perplexity-based method and generation-based method.

- **Perplexity-based Method**. On tasks that choose one single correct answer from multiple candidates such as CHID and CMRC2017, we compare the per-token perplexity score [19] when filling each answer into the blank of the context. The one with lower per-token perplexity score will be the predicted as the correct answer. On tasks that require binary or multiple classification, we assign each label with a more semantically meaningful name and use a prompt to formalize the context and the label as a human-readable text. Then, this kind of tasks can be treated as multi-choice tasks. The prompts we used are similar to that in CPM-1 and PanGu-α.

- **Generation-based Method**. On tasks with free-form completion such as Closed-book QA, we use beam search with a beam width of 8 and no length penalty. The maximum generated length of a completion is limited by a pre-defined number based on 95% percentile point of answers' length on the dataset. Then metrics such as exact match (EM), F1 and Rouge-1 are used. On tasks with restrained completion such as extractive MRC, we use restrained beam search with the same parameters as before. A Trie-Tree is constructed for each sample to efficiently and effectively restrain the space of generation and only generate completion occurred in a given text.

4.3.2 Results

Chinese News Classification. For the TNEWS and IFLYTEK datasets, there are 15 and 119 categories respectively. We randomly sample three candidates as negative labels for each sample and compare the per-token perlexity score among these four choices. This sampling strategy is aligned with CPM-1's and PanGu-α's to reduce the total computational cost since we need to calculate per-token perlexity score for each candidate separately. ERNIE 3.0 performs well on TNEWS even reaching competitiveness with prior state-of-the-art fine-tuning approaches and performs slightly well on IFLYTEK.

Semantic Similarity. We consider AFQMC and CSL datasets. ERNIE 3.0 outperforms baselines at a large margin. However, the accuracy is slightly above than a random-guess model. This may be partly attributed to the sub-optimal selection of the prompt (like THE FOLLOWING TWO SENTENCES HAVE THE SAME/DIFFERENT SEMANTICS: $SENT_A. $SENT_B.).

Natural Language Inference. ERNIE 3.0 is evaluated on two NLI datasets, namely OCNLI and CMNLI where CMNLI consists of XNLI and MNLI by translating English to Chinese. We use the prompt as $SENT_A? NO/YES/MAYBE, $SENT_B. The performance of ERNIE 3.0 is comparable to baselines, it shows that there is still a large room for improvement for pre-trained models on zero-shot NLI task.

Winograd Schema Challenge: We formalize the WSC2020 dataset as a multi-choice completion task where a pronoun is replaced with each candidates to calculate the per-token perplexity of a sample. ERNIE 3.0 improves the performance by 3.38 percent point compared to PanGu-α-13B.

---

[18] https://www.luge.ai/

[19] The perplexity score of a sample is normalized by the number of tokens.

Type	Task (# of cases)	CPM-1	PLUG	PanGu-α	ERNIE 3.0
Question Answering	Factual QA (30)	**1.67/1.50/1.03**	1.23/0.83/0.27	1.60/1.07/0.60	**1.67/1.50/1.03**
	Opinion QA (30)	1.27/0.80/-	1.43/1.13/-	1.60/1.23/-	**1.67/1.33/-**
	Reasoning (30)	1.20/0.83/**0.27**	1.03/0.83/0.07	1.03/0.83/0.00	**1.70/1.60**/0.23
Interpretation	Interpretation of Terms (30)	1.23/0.73/0.70	1.50/0.97/0.80	1.57/0.97/0.70	**1.83/1.60/1.33**
	Reverse Dictionary (30)	0.11/0.11/0.07	1/0.86/0.36	1.32/1.00/**1.00**	**1.43/1.32**/0.93
Dialogue	Single-Turn Dialogue (30)	1.63/**0.90**/-	1.37/0.17/-	1.40/0.87/-	**1.83**/0.70/-
	Multi-Turn Dialogue (50)	1.10/0.83/-	0.80/0.87/-	1.10/1.03/-	**1.43/1.33/-**
Text Generation	Recipe Generation (30)	0.80/0.63/-	**1.67**/1.03/-	1.40/1.03/-	1.30/**1.10**/-
	Novel Generation (50)	0.87/0.93/-	1.20/1.00/-	1.23/1.03/-	**1.27/1.13/-**
	Professional Manuscripts Generation (50)	0.97/0.83/-	**1.37**/1.07/-	1.23/0.83/-	1.33/**1.10/-**
	Couplet Generation (30)	0.73/0.60/-	0.77/0.86/-	1.10/0.90/-	**1.50/1.47/-**
	Poetry Generation (30)	1.80/1.60/-	1.17/1.00/-	1.833/1.07/-	**1.87/1.30/-**
Summarization	Chinese News Summarization (30)	1.21/1.10/-	0.93/0.86/-	1.24/1.03/-	**1.41/1.31/-**
Average		1.03/0.81/0.52	1.21/0.95/0.375	1.38/1.00/0.58	**1.54/1.34/0.88**

Table 6: The zero-shot generation performance manually evaluated on our collected 450 cases. (we reported the average score of *coherence*, *fluency*, and *accuracy* respectively on a scale of $[0, 1, 2]$)

Cloze and completion. On the CHID dataset, we split each sentence that contains only one blank word as a sample, and formalize as a multi-choice task. ERNIE 3.0 achieves the best score among baselines. For Chinese Word Prediction with Long Context (Chinese WPLC), a sample consists of a masked text and a correct word. Following PanGu-α, we replace the mask token with the correct word and calculate the perplexity score of a whole sentence. Compared to PanGu-α, ERNIE 3.0 achieves much lower perplexity score. On the CMRC2019 dataset, we randomly sample three negative candidates for each blank from the original candidates, then beam search is applied to calculate the optimal path for a sample. We also formalize the PD, CFT and CMRC2017 as multi-choice tasks where the text before the blank is taken as the input, and the multiple choices are the words the appear in the whole text. ERNIE 3.0 surpassed the baselines with a large margin.

Machine Reading Comprehension. We consider four MRC datasets. On C3, a multi-choice machine reading comprehension tasks, we use the prompt as QUESTION: $QUESTION? ANSWER: $CHOICE. THE ANSWER IS IN THE FOLLOWING DOCUMENT: $DOCUMENT. For CMRC2018, DRCD and DuReader, we evaluate ERNIE 3.0 using generation-base method and the prompt is DOCUMENT: $DOCUMENT. QUESTION: $QUESTION? ANSWER:. ERNIE 3.0 outperforms baselines with a large margin on CMRC2018, DRCD and DuReader dataset.

Closed-book Question Answering. We evaluated ERNIE 3.0 on two Closed-book Question Answering datasets which require the model to generate answers using its inherent knowledge learned during pre-training. WebQA is a large scale real-word QA dataset from Baidu Zhidao. We only provide ERNIE 3.0 with the question without additional evidence. The prompt is similar to MRC's but without document input (QUESTION: $QUESTION? ANSWER:). ERNIE 3.0 achieves better performance compared to baselines. We presented the detailed analysis about CKBQA dataset in Section. 5.

4.3.3 Case Study

We manually collected 450 cases to evaluate the zero-shot generation ability of current large-scale pre-trained models on 13 tasks from 5 different types including Question Answering, Interpretation, Dialogue, Text Generation and Summarization. In human evaluation, the annotators are asked to score the generation quality on a scale of $[0, 1, 2]$. We reported the average score of *coherence*, *fluency*, and *accuracy* in Tab. 6, and showed some zero-shot generations of ERNIE 3.0 in Tab. 7. ERNIE 3.0 can generate the most coherent, fluent and accurate texts on average as compared to CPM-1, PLUG, PanGu-α [20]. The introduction of three scoring metrics are listed as follows, and the scoring details are provided in Tab. 8.

- **Coherence** measures whether the generation is relevant and consistent with the context.
- **Fluency** evaluates whether the generated text is natural or readable. A fluent text should have no semantic contradiction among the generated text.

---

[20]We use the implementation of CPM-1 in `https://github.com/jm12138/CPM-Generate-Paddle`, PLUG in `https://nlp.aliyun.com/portal?/BigText_chinese#/BigText_chinese` and PanGu-α in `https://git.openi.org.cn/PCL-Platform.Intelligence/PanGu-Alpha`

Input (Factural QA)	问题:发明加减消元法的数学家所在国家有哪些品牌汽车？ 答:
Generation	有很多，比如说奔驰、宝马、奥迪、大众、保时捷。
Input (Opinion QA)	问题:宝宝眼睛流泪是怎么回事？ 答:
Generation	这种情况可能有三个原因：一是宝宝泪腺的炎症。二是宝宝的眼睛有倒睫毛或者睫毛刺激眼球引起流泪。三是宝宝有结膜炎。所以如果宝宝眼睛流泪，首先要排除宝宝倒睫毛刺激眼球引起的流泪，如果不是倒睫毛引起的流泪，还要考虑有没有泪腺炎症或者结膜炎的可能。
Input (Reverse Dictonary)	问题:描写"人的品质优良"的词语有哪些？ 答:
Generation	品质优良,品德高尚,诚实,忠厚,老实
Input (Novel Generation)	江西孟龙潭与朱孝廉客都中，偶涉一兰若，殿宇禅舍，俱不甚弘敞，惟一老僧挂褡其中。见客入，肃衣出迓，导与随喜。殿中塑志公像，两壁画绘精妙，人物如生。东壁画散花天女，内一垂髫者，拈花微笑，樱唇欲动，眼波将流。
Generation	西壁画一小沙弥，面目奇古，双手合十，似在祈祷。客大骇，请问之。僧曰："此乃佛经也。"客曰："此何所谓？"僧曰："此有佛性，无所不有。"客曰："何以见之？"僧曰："佛性即是佛性。佛性无所不有，所以不能言说。"客大悟，遂去。
Input (Couplet Generation)	上联:口十心思，思乡，思友，思父母 下联:
Generation	眼四面观，观景，观人，观自己
Input (News Summarization)	花盆里有黑色小飞虫，这应该是夏天大家都比较头疼的问题。灭小飞虫方法很多，给大家分享几种。电蚊拍电，粘虫板粘，杀虫剂喷，烟蒂水灌根，诱杀法，其他异味水灌根。消灭花盆里的小黑飞虫，可以把烟蒂水灌根和电蚊拍拍打结合起来，坚持一段时间就能彻底消灭。大蒜香烟泡一泡，用这一碗水，小虫去无踪。文章标题是《
Generation	花盆里有黑色小飞虫怎么办？ 》

Table 7: Illustrations of zero-shot generations from ERNIE 3.0.

Score	Coherence	Fluency	Accuracy
0	The generation is not related to the context. The generation has obvious conflicts with the context.	The generation is unnatural. There are contradictions in the generated text.	The answer is wrong.
1	The generation is weakly related to the context. The generation has minor logic conflicts with the context.	The generation has minor influent part. The generation slightly influences the reading.	The answer is partly correct.
2	The generation is strongly related to the context. The logic in the generation is aligned with the context.	The generation is semantically complete and fluent. There are no contradictions in the generated text.	The answer is correct.

Table 8: Scoring details for zero-shot generation.

- **Accuracy** is a metric to evaluate whether the generated text is the same as the ground truth.

4.4 Experiments on SuperGLUE

Model	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	Score
Human Baseline	89.0	95.8/98.9	100	81.8/51.9	91.7/91.3	93.6	80.0	100	89.8
T5+Menna	**91.4**	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	**77.9**	96.6	90.4
DeBERTa	90.4	95.7/97.6	**98.4**	88.2/**63.7**	94.5/94.1	**93.2**	77.5	95.9	90.3
ERNIE 3.0	91.0	**98.6/99.2**	97.4	**88.6**/63.2	**94.7/94.2**	92.6	77.4	**97.3**	**90.6**

Table 9: SuperGLUE test set results which are scored by the SuperGLUE evaluation server (Results are recorded at July 3, 2021 from https://super.gluebenchmark.com/leaderboard).

As a multi-task benchmark for natural language understanding, SuperGLUE [3] is usually used to evaluate the performance of pre-training models. We also test the performance of ERNIE 3.0 on SuperGLUE, which covers a diverse range of NLP datasets as follows.

- BoolQ (Boolean Questions, [56]) is a QA task where each example consists of a short passage and a yes/no question about the passage. The task is to answer the questions with YES or NO, and the metric of this task is accuracy.

- CB (Commitment Bank, [88]) is an imbalanced corpus of natural language inference task. The task is evaluated using accuracy and macro-F1.

- COPA (Choice of Plausible Alternatives [89]) is a causal reasoning task based on common sense knowledge. The data are curated from blogs and a photography-related encyclopedia. Following the original work, we evaluate this task using accuracy.

- MultiRC (Multi-Sentence Reading Comprehension [90]) is a QA task where each example consists of a context paragraph, a question about that paragraph, and a list of possible answers. The system must predict which answers are true and which are false. The evaluation metrics are F1 over all answer-options ($F1_a$) and exact match of each question's set of answers (EM).

- ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset, [91]) is a multiple-choice QA task. It requires the model to pick an entity to complete the answer, given a context of news article and a Cloze-style question. This task is evaluated with max (over all mentions) token-level F1 and exact match.

- RTE (Recognizing Textual Entailment [92]) dataset comes from a series of annual competitions on textual entailment. It is a natural language inference corpus and evaluated with accuracy.

- WiC (Word-in-Context [93]) is a word sense disambiguation task cast as binary classification of sentence pairs using accuracy as the evaluation metrics.

- WSC (Winograd Schema Challenge [94]) is a coreference resolution task in which examples consist of a sentence with a pronoun and a list of noun phrases from the sentence as choices. The system must select the correct referent of the pronoun from the provided choices. This task is evaluated with accuracy.

Similar to the pre-training corpus used in RoBERTa [95] and DeBERTa [96], we compiled the English pre-training corpus for ERNIE 3.0 including English Wikipedia, BookCorpus [97], CC-News [98], OpenWebText [99], Stories [100]. As shown in the Table 9, ERNIE 3.0 surpasses T5 [1] and DeBERTa [96] and obtains a score of 90.6, taking the first place in SuperGLUE Benchmark.

5 Analysis

The Effectiveness of the Task-specific Representation Modules
To verify the effectiveness of the task-specific networks, we compare our proposed structure with those which share parameters under various pre-training tasks. For the ablation test, we choose understanding and generation as two different training paradigms and utilize the corresponding tasks mentioned in Section 3.2. The unified network follows the base model settings (12 layers, 768 dims, 12 attention heads), and the task-specific networks for each task paradigms are set to 3 layers, 256 dims, and 4 attention heads. For the contrast model, the task-specific network is shared across different task paradigms. Figure 3 illustrates the perplexity variation of the NLG task during the pre-training process.

As shown in Figure 3, the model with its own task-specific network for different task paradigms reaches a higher convergence speed. Furthermore, as training progresses, the performance gap becomes bigger compared to the model with a shared task-specific network. The experimental result shows the effectiveness of the proposed task-specific networks and demonstrates the necessity of distinguishing different tasks.
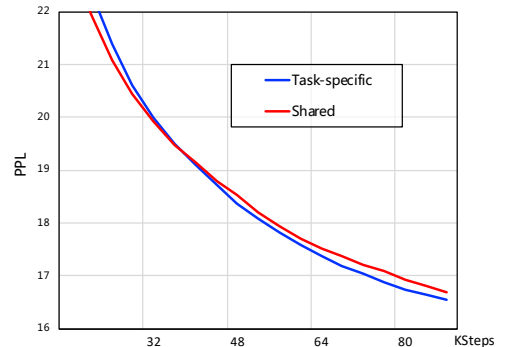

Figure 3: Perplexity variation of the NLG pre-training task with respect to training steps.

Universal Knowledge-Text Prediction   A group of ablation experiments is conducted to evaluate the performance of the universal knowledge-text prediction task. The relation extraction task is a typical knowledge-driven task, aiming to predict the relationship between two entities mentioned in a given sentence. Specifically, we add four special tokens, [HD], [/HD], [TL] and [/TL] to identify the mention of a head entity and a tail entity respectively, then the relation classification is performed on the sum of the final representations of the aforementioned four special tokens. We

construct the experiments on SanWen and FinRE datasets and as shown in Table 10, the knowledge enhancement strategy achieves impressive empirical performance on the relation extraction task.

In addition, the zero-shot generation experiment on CKBQA also confirms the effectiveness of the universal knowledge-text prediction task. Specifically, the knowledge-based question answering (KBQA) task requires a model to search and reason for correct answers based on a knowledge graph. It's suitable to measure the knowledge learning capability of the pre-trained languages models using the KBQA task. We use the "QUESTION: $QUESTION? ANSWER:" as the prompt for zero-shot learning and then compare the performance of our proposed model with several state-of-the-art pre-trained language models

Dataset	ERNIE$_{\text{Base}}$	ERNIE$_{\text{Base}}$+UKTP
SanWen	75.56	77.36(+1.80)
FinRE	58.19	59.75(+1.56)

Table 10: Ablation experiments of universal knowledge-text prediction task.

on the CKBQA dataset. As shown in Table 5, ERNIE 3.0 significantly outperforms PanGu-α and CPM-1 in the CKBQA dataset which indicates that ERNIE 3.0 has the ability to memorize and learn more knowledge.

Progressive Learning to Speed up Convergence We record the training convergence speed on two architecture settings including ERNIE$_{\text{Base}}$ and ERNIE$_{\text{1.5B}}$, in which the architecture settings of ERNIE$_{\text{Base}}$ follows [7] and ERNIE$_{\text{1.5B}}$ model consists of 48 layers with a hidden size of 1,536 and 24 attention heads. As shown in Tab. 11, we record the time for the loss value of the model converges to the same as that of the ERNIE 3.0. For the ERNIE$_{\text{Base}}$ model, the convergence time is reduced by 65.21% from 11 hours to 4 hours, and for the ERNIE$_{\text{1.5B}}$, the convergence time is reduced by 48%. For both two settings, we carry out pre-training on 8 NVIDIA Tesla V100 GPUs. For ERNIE$_{\text{Base}}$, we increased the batch size from 8 to 2048 and the sequence length from 128 to 512, the learning rate increases linearly from 0 to 1e-4, and the dropout keeps 0 in the progressive warmup stage. For

Method	Training Time
ERNIE$_{\text{Base}}$ +Progressive Learning	11h30m 4h(-65.21%)
ERNIE$_{\text{1.5B}}$ +Progressive Learning	5h55m 3h4m(-48.2%)

Table 11: Progressive Learning To Speedup Training.

ERNIE$_{\text{1.5B}}$, we gradually increase the batch size from 8 to 8192, The learning rate increases from 0 to 6e-4, the dropout also keeps 0. The rest settings for the experiment remain as same as [7]. For ERNIE$_{\text{1.5B}}$, to achieve the peak batch size within the constraint of GPU memory, the gradient accumulation strategy is used during the pre-training.

6 Conclusion

we proposed the ERNIE 3.0 framework to pre-train a knowledge enhanced 10-billion parameter model on a 4TB corpus including plain texts and a knowledge graph. In order to handle both language understanding and generation tasks with zero-shot learning, few-shot learning and fine-tuning, ERNIE 3.0 designs a unified pre-training framework that integrates both auto-encoder networks and auto-regressive networks. We construct extensive experiments on various datasets from different task paradigms and fields, and the results demonstrate the effectiveness of ERNIE 3.0 as compared to the previous state-of-the-art pre-trained models.

References

[1] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[3] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*, 2019.

[4] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*, 2018.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] 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. *arXiv preprint arXiv:1904.09223*, 2019.

[8] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[9] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[10] Baotian Hu, Qingcai Chen, and Fangze Zhu. Lcsts: A large scale chinese short text summarization dataset. *arXiv preprint arXiv:1506.05865*, 2015.

[11] Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

[12] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *CoRR*, abs/1909.08053, 2019.

[13] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *CoRR*, abs/2101.03961, 2021.

[14] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

[16] Matthew Hutson. Robo-writers: the rise and risks of language-generating ai. Website, 2021. `https://www.nature.com/articles/d41586-021-00530-0`.

[17] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[18] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

[19] Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, et al. Cpm: A large-scale generative chinese pre-trained language model. *arXiv preprint arXiv:2012.00413*, 2020.

[20] Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, et al. Cpm-2: Large-scale cost-effective pre-trained language models. *arXiv preprint arXiv:2106.10715*, 2021.

[21] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823*, 2021.

[22] Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. Pangu-$alpha$: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*, 2021.

[23] Hyperclova from naver. `https://medium.com/ai-trend/if-you-look-at-the-direction-of-naver-ai-you-can-feel-the-potential-of-ai-network-bb129aa9b73a`.

[24] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*, 2019.

[25] Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*, 2019.

[26] Bin He, Di Zhou, Jinghui Xiao, Qun Liu, Nicholas Jing Yuan, Tong Xu, et al. Integrating graph contextualized knowledge into pre-trained language models. *arXiv preprint arXiv:1912.00147*, 2019.

[27] Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. *arXiv preprint arXiv:1912.09637*, 2019.

[28] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021.

[29] Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. Colake: Contextualized language and knowledge embedding. *arXiv preprint arXiv:2010.00309*, 2020.

[30] Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, Bill Yuchen Lin, and Xiang Ren. Pre-training text-to-text transformers for concept-centric common sense. *arXiv preprint arXiv:2011.07956*, 2020.

[31] 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.

[32] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[33] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975, 2020.

[34] Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[35] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[36] He Bai, Peng Shi, Jimmy Lin, Luchen Tan, Kun Xiong, Wen Gao, and Ming Li. Segabert: Pre-training of segment-aware BERT for language understanding. *CoRR*, abs/2004.14996, 2020.

[37] Siyu Ding, Junyuan Shang, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-doc: The retrospective long-document modeling transformer. *arXiv preprint arXiv:2012.15688*, 2020.

[38] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[39] Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation, 2020.

[40] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[42] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training. *CoRR*, abs/2104.00298, 2021.

[43] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[44] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ICLR*, 2018.

[45] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*, 2020.

[46] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[47] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[48] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.

[49] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[50] Yanzeng Li, Tingwen Liu, Diying Li, Quangang Li, Jinqiao Shi, and Yanqiu Wang. Character-based bilstm-crf incorporating pos and dictionaries for chinese opinion target extraction. In *Asian Conference on Machine Learning*, pages 518–533. PMLR, 2018.

[51] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

[52] Ziran Li, Ning Ding, Zhiyuan Liu, Haitao Zheng, and Ying Shen. Chinese relation extraction with multi-grained information and external linguistic knowledge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4377–4386, 2019.

[53] Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. *arXiv preprint arXiv:1711.07010*, 2017.

[54] Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. Lcqmc: A large-scale chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1952–1962, 2018.

[55] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*, 2019.

[56] Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang. The bq corpus: A large-scale domain-specific chinese corpus for sentence semantic equivalence identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4946–4951, 2018.

[57] LTD IFLYTEK CO. Iflytek: a multiple categories chinese text classifier. competition official website, 2019.

[58] Bang Liu, Di Niu, Haojie Wei, Jinghong Lin, Yancheng He, Kunfeng Lai, and Yu Xu. Matching article pairs with graphical decomposition and convolutions. *arXiv preprint arXiv:1802.07459*, 2018.

[59] Sheng Zhang, Xin Zhang, Hui Wang, Jiajun Cheng, Pei Li, and Zhaoyun Ding. Chinese medical question answer matching using end-to-end character-level multi-scale cnns. *Applied Sciences*, 7(8):767, 2017.

[60] Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071, 2018.

[61] Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. *arXiv preprint arXiv:1607.06275*, 2016.

[62] Nanyun Peng and Mark Dredze. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, 2015.

[63] Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, 2011.

[64] Yiming Cui, Ting Liu, Li Xiao, Zhipeng Chen, Wentao Ma, Wanxiang Che, Shijin Wang, and Guoping Hu. A span-extraction dataset for chinese machine reading comprehension. *CoRR*, abs/1810.07366, 2018.

[65] Yiming Cui, Ting Liu, Ziqing Yang, Zhipeng Chen, Wentao Ma, Wanxiang Che, Shijin Wang, and Guoping Hu. A sentence cloze dataset for chinese machine reading comprehension. *arXiv preprint arXiv:2004.03116*, 2020.

[66] Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. Drcd: a chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*, 2018.

[67] Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*, 2017.

[68] Hongxuan Tang, Jing Liu, Hongyu Li, Yu Hong, Hua Wu, and Haifeng Wang. Dureaderrobust: A chinese dataset towards evaluating the robustness of machine reading comprehension models. *arXiv preprint arXiv:2004.11142*, 2020.

[69] Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Investigating prior knowledge for challenging chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155, 2020.

[70] Chujie Zheng, Minlie Huang, and Aixin Sun. Chid: A large-scale chinese idiom dataset for cloze test. *arXiv preprint arXiv:1906.01265*, 2019.

[71] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*, 2018.

[72] Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. Blow the dog whistle: A Chinese dataset for cant understanding with common sense and world knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2139–2145, Online, June 2021. Association for Computational Linguistics.

[73] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, pages 55–64, 2017.

[74] Canwen Xu, Jiaxin Pei, Hongtao Wu, Yiyu Liu, and Chenliang Li. Matinf: A jointly labeled large-scale dataset for classification, question answering and summarization. *arXiv preprint arXiv:2004.12302*, 2020.

[75] Yan Wang, Xiaojiang Liu, and Shuming Shi. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 845–854, 2017.

[76] Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. Long and diverse text generation with planning-based hierarchical variational model. *arXiv preprint arXiv:1908.06605*, 2019.

[77] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, et al. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, 2020.

[78] Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. Kdconv: a chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. *arXiv preprint arXiv:2004.04100*, 2020.

[79] Dongling Xiao, Yu-Kun Li, Han Zhang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-gram: Pre-training with explicitly n-gram masked language modeling for natural language understanding. *arXiv preprint arXiv:2010.12148*, 2020.

[80] Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635*, 2020.

[81] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*, 2019.

[82] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[83] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*, 2020.

[84] Yan Song, Tong Zhang, Yonggang Wang, and Kai-Fu Lee. Zen 2.0: Continue training and adaption for n-gram enhanced text encoders. *arXiv preprint arXiv:2105.01279*, 2021.

[85] Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. Glyce: Glyph-vectors for chinese character representations. *arXiv preprint arXiv:1901.10125*, 2019.

[86] Xiongtao Cui and Jungang Han. Chinese medical question answer matching based on interactive sentence representation learning. *arXiv preprint arXiv:2011.13573*, 2020.

[87] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[88] Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. 2019. To appear in proceedings of Sinn und Bedeutung 23. Data can be found at https://github.com/mcdm/CommitmentBank/.

[89] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011.

[90] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, 2018.

[91] Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint 1810.12885*, 2018.

[92] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer, 2006.

[93] Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*, 2019.

[94] Hector J Levesque, Ernest Davis, and Leora Morgenstern. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47, 2011.

[95] 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, 2019. cite arxiv:1907.11692.

[96] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2020.

[97] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

[98] Sebastian nagel. 2016. cc-news. `http://web.archive.org/save/http://commoncrawl.org/2016/10/news-dataset-available`.

[99] Aaron gokaslan and vanya cohen. 2019. openweb-text corpus. `http://web.archive.org/save/http://Skylion007.github.io/OpenWebTextCorpus`.

[100] Trieu H Trinh and Quoc V Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.