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

Yu Sun et al.

집현전 최신반

2021-08-15

김은희, 박한솔, 구혜연

1. Ernie 3.0에 대한 소개

1 현재 SuperGlue SOTA 모델

- Human Super Glue 89.8 점을 넘은 모델이 Deberta, T5+ Meena에 이어 3번째 모델
- 10 billion 학습 parameter들을 plain texts와 large scale knowledge corpus로 구성된 4TB dataset으로 실험

2 Ernie3.0 모델의 특징

- Ernie 3.0은 입력 데이터에 대해 Universal Representation Module 이후 Task Specific Representation Module로 구성된 Continual Multi-Paradigms Unified Pre-training Framework을 제시
- Task Specific Representation Module 은 Language Understanding representation과 Language Generation representation 두 모듈로 구성됨

INDEX

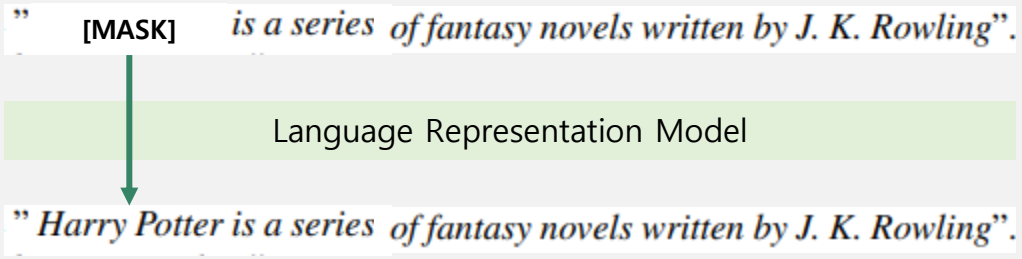
01. Introduction

1. Introduction

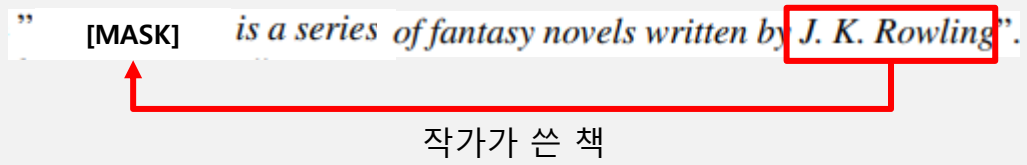
”Harry Potter is a series of fantasy novels written by J. K. Rowling”.

소설 이름 작가

- 모델이 긴 컨텍스트의 도움 없이 내부의 단어 배열로 누락된 단어를 예측하는 것은 쉬움



- 그러나, 모델은 **엔티티 사이의 관계를 이용해서 누락된 단어를 예측할 수 없음**



작가가 쓴 책

ERNIE:
enhanced representation through **Knowledge** integration

1 새로운 학습 절차

- knowledge masking : 몇 개의 단어가 포함된 phrase, entity를 하나의 unit으로 보고, unit에 포함된 모든 단어들을 masking
- knowledge embedding을 직접 추가하는 대신에, ERNIE는 knowledge 정보와 긴 semantic 의존성을 학습할 수 있도록 단어 embedding 학습을 유도함

2 성능 향상

- 이전에 다양한 중국어 NLP task에서 SOTA였던 모델들보다 성능이 상당히 향상됨

3 모델 및 코드 공개

- https://github.com/PaddlePaddle/ERNIE

02. Methods

1. Transformer Encoder & Knowledge Integration

1 Transformer Encoder

- ERNIE는 multi-layer Transformer를 기본 **encoder**로 사용
  - Self-attention을 통해 contextual information 포착
- 중국어 corpus를 전처리
  - **CJK 유니코드 범위의 모든 문자 주위에 공백 추가**
  - **WordPiece를 tokenizer**로 사용
  - token, segment, position embedding를 input representation으로 사용
  - 모든 sequence 맨 앞에 [CLS] 추가

2 Knowledge Integration

- PLM을 향상시키기 위해 사전(prior) 지식을 사용함
- knowledge embedding을 직접 추가하지 않고, **multi-stage knowledge masking**을 통해 phrase, entity level의 knowledge를 integration함
- 문장에서 각 level마다 다른 masking을 적용 : Figure 2

Sentence	Harry	Potter	is	a	series	of	fantasy	novels	written	by	British	author	J.	K.	Rowling
Basic-level Masking	[mask]	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	J.	[mask]	Rowling
Entity-level Masking	Harry	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]
Phrase-level Masking	Harry	Potter	is	[mask]	[mask]	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]

entity masking

phrase masking

Figure 2: Different masking level of a sentence

03. Experiments

1. Corpus & DLM

1 Heterogeneous Corpus Pre-training

- pre-training을 위해
  이질적인 corpus를 섞어서 사용
  - Wikepedia : 21M(단위 : 문장)
  - Baidu Baike : 51M
    formal language로 쓰인 백과사전
  - Baidu News: 47M
    영화, 배우, 팀 이름 등 최신 정보
  - Baidu Tiebe: 54M
    open discussion forum으로
    각 post는 dialog thread가 됨
- 중국어는 traditional-to-simplified로
  대화를 가공하고,
  영어는 upper-to-lower로 가공
- shared-vocabular로 17,964개의
  Unicode 문자를 사용

2 DLM (Dialogue Language Model)

- ERNIE는 Query-Response 대화 구조를 모델링함
- ERNIE의 Dialogue embedding은 multi-turn 대화를 나타낼 수 있음 (QRQ, QRR, QQR 등)
- query, response 모두에서 masking된 단어를 예측하도록 강제함
- random하게 선택된 문장에서 query 또는 response를 바꿔 fake sample을 생성
- **DLM Task는 ERNIE가 dialogue에서 함축적인 관계를 학습할 수 있게 하고, semantic representation을 학습하는 능력을 향상시킴**

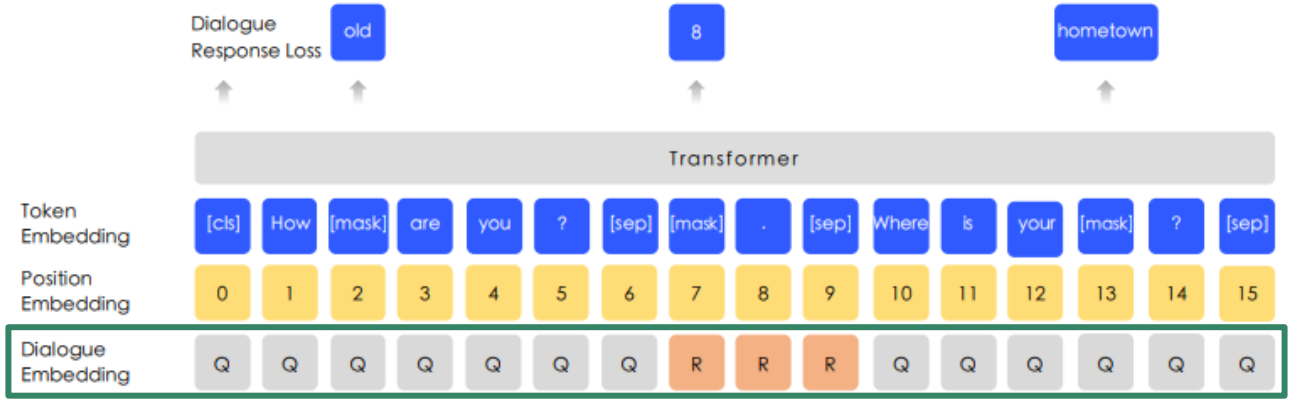


Figure 3: Dialogue Language Model. Source sentence: [cls] How [mask] are you [sep] 8 . [sep] Where is your [mask] ? [sep]. Target sentence (words the predict): old, 8, hometown)

2. Experiments & Results on Chinese NLP tasks

1 실험 데이터셋

Task	Dataset	Description
NLI	XNLI (Cross-lingual Natural Language Inference)	• 언어 관계가 labeling 되고 14개 언어로 번역됨 • contradiction, neutral, entailment(모순, 중립, 포함
Semantic Similarity	LCQMC (Large-scale Chinese Question Matching Corpus)	• 두 문장이 동일한 의도를 가졌는지 판단 • binary label로 정형화됨
NER	MSRA-NER	• Microsoft Research Asia 에서 만듦 • 인명, 지명, 기관명 등을 포함
Sentiment Analysis	ChnSentiCorp	• 문장의 감정을 판단 • positive, negative
Retrieval Question Answering	NLPCC-DBQA	• 해당 질문에 대한 답변 찾기 • 평가지표에 MRR과 F1 score가 포함됨

2 실험 결과

Table 1: Results on 5 major Chinese NLP tasks

Task	Metrics	Bert		ERNIE	
		dev	test	dev	test
XNLI	accuracy	78.1	77.2	79.9 (+1.8)	78.4 (+1.2)
LCQMC	accuracy	88.8	87.0	89.7 (+0.9)	87.4 (+0.4)
MSRA-NER	F1	94.0	92.6	95.0 (+1.0)	93.8 (+1.2)
ChnSentiCorp	accuracy	94.6	94.3	95.2 (+0.6)	95.4 (+1.1)
nlpcc-dbqa	mrr	94.7	94.6	95.0 (+0.3)	95.1 (+0.5)
	F1	80.7	80.8	82.3 (+1.6)	82.7 (+1.9)

- **ERNIE는 모든 task에서 BERT보다 좋은 성능을 보임**
- XNLI, MSRA-NER, ChnSentiCorp, nlpcc-dbqa에서는 BERT보다 1% 이상의 성능 향상을 보임
- ERNIE는 knowledge integration 전략으로 더 좋은 점수를 얻음

3. Ablation Studies

1 Effect of Knowledge Masking Strategies

Table 2: XNLI performance with different masking strategy and dataset size

pre-train dataset size	mask strategy	dev Accuracy	test Accuracy
10% of all	word-level(chinese character)	77.7%	76.8%
10% of all	word-level&phrase-level	78.3%	77.3%
10% of all	word-level&phrase-leve&entity-level	78.7%	77.6%
all	word-level&phrase-level&entity-level	79.9 %	78.4%

- **phrase-level** mask를 추가하는 것이 모델의 **성능을 향상**시킬 수 있음
- **entity-level** masking 전략으로 **성능을 한 층 더 향상**시킬 수 있음
- 전체 데이터셋으로 실험했을 때, XNLI 데이터셋에 대해서 총 0.8%의 성능 향상이 있었음

2 Effect of DLM (Dialogue Language Model)

Table 3: XNLI finetuning performance with DLM

corpus proportion(10% of all training data)	dev Accuracy	test Accuracy
Baike(100%)	76.5%	75.9%
Baike(84%) / news(16%)	77.0%	75.8%
Baike(71.2%)/ news(13%)/ forum Dialogue(15.7%)	77.7%	76.8%

※ Baike : formal language로 쓰인 백과사전
News : 영화배우, 팀 이름 등 최신 정보
Tiebe: open discussion forum, 각 post가 dialogue thread가 됨

- ERNIE를 이 데이터셋에 대해 pre-training하고, random으로 5개의 fine-tuning한 결과의 평균
- DLM Task가 포함되었을 때, develop 0.7%, test 1.0%의 성능 향상이 있었음

4. Cloze Test

1 빈 칸 채우기 실험

- case1 : **BERT**는 **context에 등장하는 name을 복사**하는 반면, **ERNIE는 article에 언급된 관계에 대한 지식을 기억함**
- case2, 5 : **BERT**는 context에 따른 pattern을 성공적으로 학습해 **named entity type은 잘 예측하지만 slot을 채우는 것에는 실패**한 반면, **ERNIE는 slot도 잘 채움**
- case3, 4, 6 : **BERT**는 문장과 관련된 몇 개의 문자로 빈 칸을 채우지만 **semantic concept를 예측하기 어려운** 반면, **ERNIE**는 case4를 제외하고 **정답을 맞춤**

No	Text	Predict by ERNIE	Predict by BERT	Answer
1	2006年9月，_____与张柏芝结婚，两人婚后育有两儿子——大儿子Lucas谢振轩，小儿子Quintus谢振南；	谢霆锋	谢振轩	谢霆锋
	In September 2006, _____ married Cecilia Cheung. They had two sons, the older one is Zhenxuan Xie and the younger one is Zhennan Xie.	Tingfeng Xie	Zhenxuan Xie	Tingfeng Xie
2	戊戌变法，又称百日维新，是_____、梁启超等维新派人士通过光绪帝进行的一场资产阶级改良。	康有为	孙世昌	康有为
	The Reform Movement of 1898, also known as the Hundred-Day Reform, was a bourgeois reform carried out by the reformists such as ____ and Qichao Liang through Emperor Guangxu.	Youwei Kang	Shichang Sun	Youwei Kang
3	高血糖则是由于_____分泌缺陷或其生物作用受损，或两者兼有引起。糖尿病时长期存在的高血糖，导致各种组织，特别是眼、肾、心脏、血管、神经的慢性损害、功能障碍。	胰岛素	糖糖内	胰岛素
	Hyperglycemia is caused by defective _____ secretion or impaired biological function, or both. Long-term hyperglycemia in diabetes leads to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves.	Insulin	(Not a word in Chinese)	Insulin
4	澳大利亚是一个高度发达的资本主义国家，首都为_____。作为南半球经济最发达的国家和全球第12大经济体、全球第四大农产品出口国，其也是多种矿产出口量全球第一的国家。	墨尔本	墨悉本	堪培拉
	Australia is a highly developed capitalist country with _____ as its capital. As the most developed country in the Southern Hemisphere, the 12th largest economy in the world and the fourth largest exporter of agricultural products in the world, it is also the world's largest exporter of various minerals.	Melbourne	(Not a city name)	Canberra (the capital of Australia)
5	_____是中国神魔小说的经典之作，达到了古代长篇浪漫主义小说的巅峰，与《三国演义》《水浒传》《红楼梦》并称为中国古典四大名著。	西游记	《小》	西游记
	_____is a classic novel of Chinese gods and demons, which reaching the peak of ancient Romantic novels. It is also known as the four classical works of China with Romance of the Three Kingdoms, Water Margin and Dream of Red Mansions.	The Journey to the West	(Not a word in Chinese)	The Journey to the West
6	相对论是关于时空和引力的理论，主要由_____创立。	爱因斯坦	卡尔斯所	爱因斯坦
	Relativity is a theory about space-time and gravity, which was founded by _____.	Einstein	(Not a word in Chinese)	Einstein

Figure 4: Cloze test

문장 내 단어 복사

사람 이름이지만 정답은 아님

문맥에 어울리지 않음

ERNIE도 틀림

문맥에 어울리지 않음

문맥에 어울리지 않음

INDEX

01. Introduction

1. ERNIE 2.0 소개

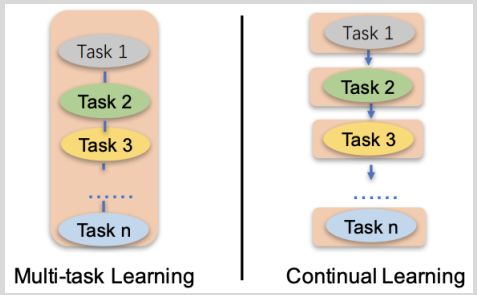
- 주로 Pre-training 과정은 단어, 문장의 co-occurring 정보 기반으로 만들어지나
  lexical, syntactic, semantic information 등의 정보가 언어를 이해하는 데에 더욱 도움이 될 것
- 이를 위해 **다양한 Taks들을 효율적으로 학습 시키는 Continual multi-task training framework, ERNIE 2.0을 소개**

1

최근 LM들은 단어, 문장의 co-occurring 정보를 기반해 문맥중심의 언어 표현을 학습하는 방향으로 발전해왔으나 ,

lexical, syntactic, semantic information 등 언어 이해에 더 중요한 정보들이 존재

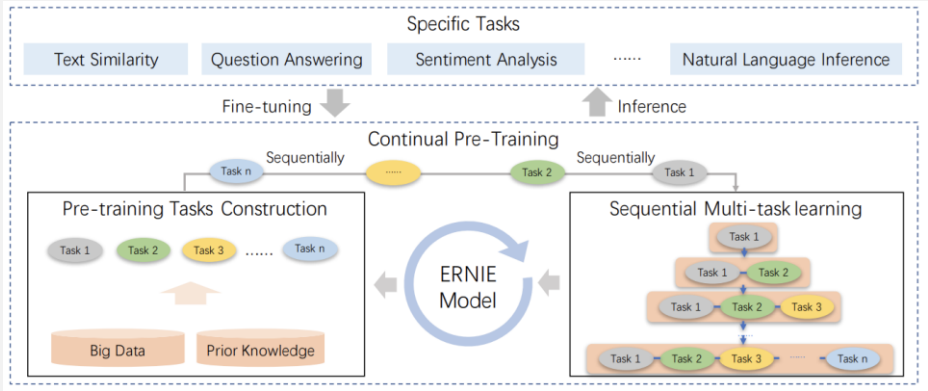
2



다양한 task를 하나의 모델에 학습시키기 위해

1) 한번에 모든 sub-task를 같이 학습하는 방법이 있지만 **새로운 task를 추가학습하기 힘들고**

2) task를 순차적으로 학습시키는 continual learning 방법이 있지만 **따로 학습시켰을 때 이전 정보학습 내용이 소실**되어 성능이 유지되지 않는 단점이 있음



1

ERNIE 2.0는 사전학습 단계에서 광범위한 Task들을 사용하여 **모델이 효율적으로 어휘적, 문맥적, 의미적 표현을 학습 할 수 있도록**함

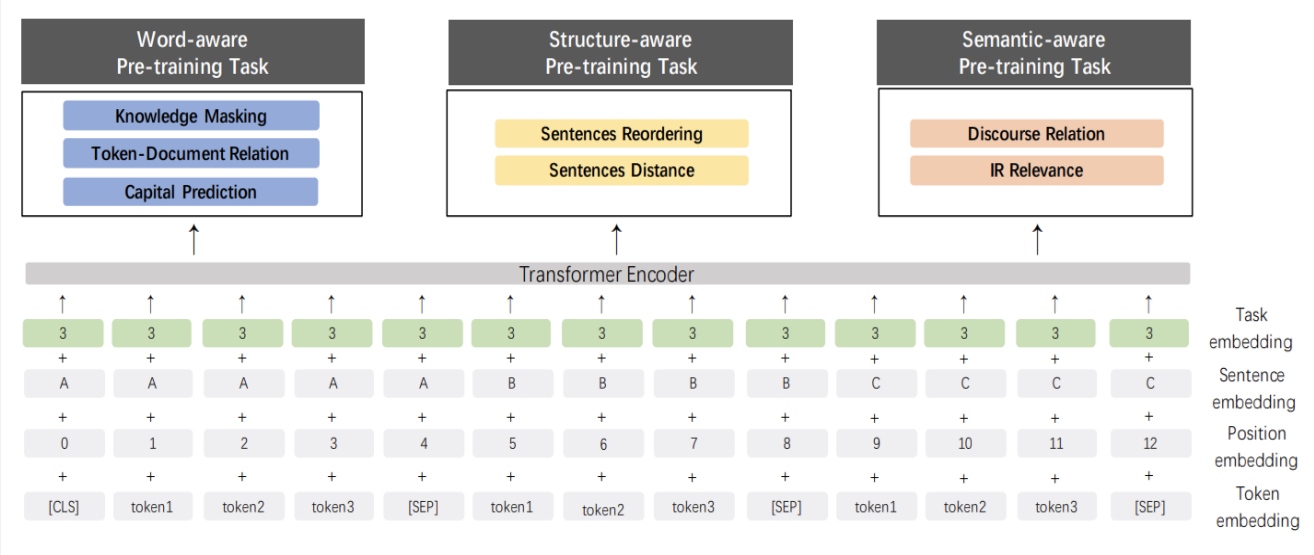
2

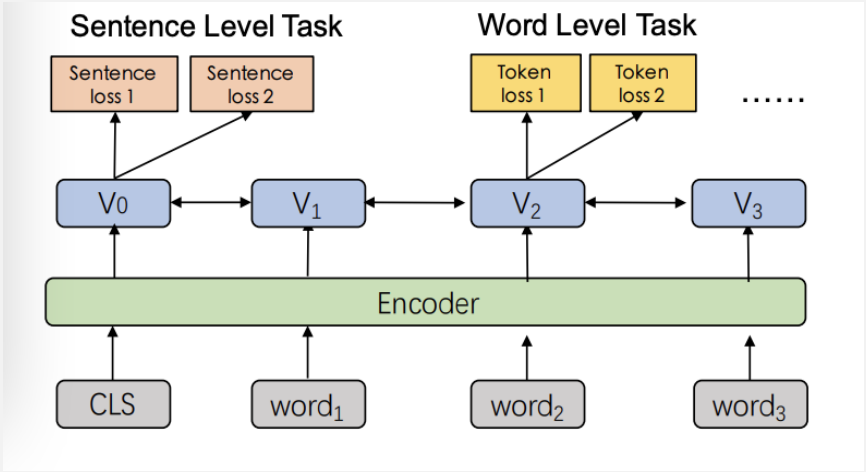
사람이 학습 할 때 지식을 순차적으로 학습한다는 사실에 착안해 **여러 task를 점진적으로 같이 학습시켜 효과적으로 언어정보 습득하며 과거 task학습내용 소실 방지**

02. Model Architecture

1. Architecture

- **3가지 사전학습 Task**를 구성하여 학습 데이터로부터 서로 다른 정보를 가져오는 것이 주된 목표
- **단어중심 Task**는 모델이 **어휘적 정보**, **문장중심 Task**에서는 **문장구조 정보**, **의미중심 Task**는 **의미정보**를 가져오는 것을 목적으로 구성





- Token Embedding, Position embedding, Sentence embedding은 기존 LM의 입력값의 형태와 동일.
  Task embedding의 경우 Task가 N개 있다면 **각각 1~N 의 id**를 갖게함
- Transfomer 를 기본 encoder로 사용, [CLS]는 시퀀스의 맨 처음에 추가로 넣고 문장 구분자로 [SEP] 추가
- 텍스트 encoding할 때는 **같은 Transformer layer를 공유**하며
  문장수준 Task에서는 **첫번째 토큰 [CLS]의 decoding 결과값**을, **단어단위 Task**에서는 **각 토큰결과값**을 이용해 **loss 구함**
- 각 사전 Task는 고유한 손실함수를 갖으며 정의된 모든 task의 loss를 최소화하는 방향으로 transformer layer는 학습됨

2. Multi-Task 종류와 기대효과

1 Word-aware Pre-training Tasks

Knowledge Masking Task

- ERNIE 1.0에서 제시한 Phrase와 Name Entity를 예측하는 task를 도입해 지역적, 전반적 문맥에서 의존성 학습

Capitalization Prediction Task

- Token이 대문자인지 소문자인지 예측하는 문제
- 대문자로 쓰여진 단어는 특별한 의미를 가지고 있는 경우가 많음.
  이 Task는 향후 NER과 같은 학습에도 도움됨

Token-Document Relation Prediction Task

- 문서에 자주 등장하는 단어를 판별하는 문제
- 어느 segment의 token이 다른 segment에 나타나는지 예측

2 Structure-aware Pre-training Tasks

Sentence Reordering Task

- 주어진 문서를 조합으로 섞은 후 segment들의 순서를 찾는 문제
- 문장 재배치는 모델이 문장관계를 학습하게 함

Sentence Distance Task

- 문장간 거리를 학습하는 Task
- 기준에 따라 class로 나눔

  0 - 두 문장이 같은 문서에 인접할 때

  1 - 두 문장이 같은 문서에 있을 때

  2 - 서로 다른 문서에 존재할 때

3 Semantic-aware Pre-training Tasks

Discourse Relation Task

- 문장 간의 담화 관계를 통해 의미 인식 표현을 추론하는 Task

IR Relevance Task

- 검색엔진에서 얻은 검색기록데이터를 사용해 텍스트 관계 분류 문제
- 분류문제로 Query, title간 관계를 예측

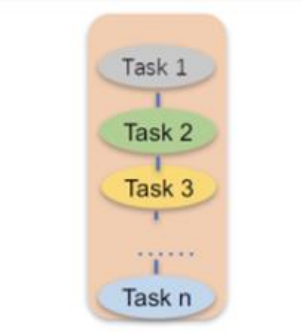
  0 - 유저가 검색 후 제목을 클릭

  1 - 제목이 검색결과로 나왔지만 클릭하지 않음

  2 - Query와 Title이 무관함

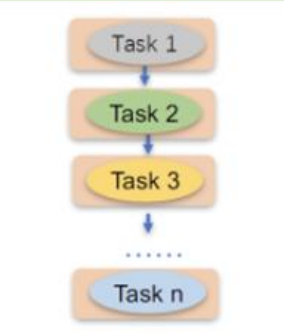
2. Continual Multi-task Learning

Multi-task Learning



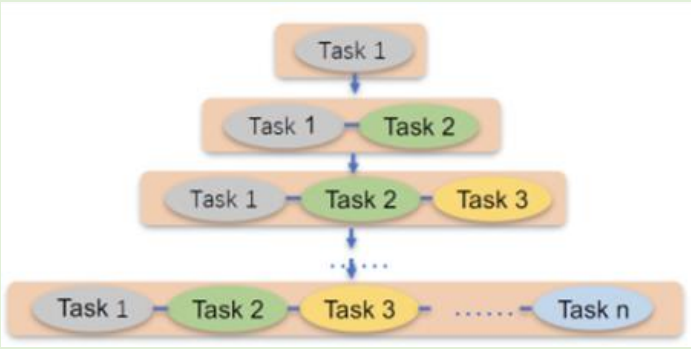
(STEP)	Stage
Task 1	50K
Task 2	50K
Task 3	50K

Continual Learning



(STEP)	Stage 1	Stage 2	Stage 3
Task 1	50K		
Task 2		50K	
Task 3			50K

Continual Multi-task Learning



(STEP)	Stage 1	Stage 2	Stage 3
Task 1	30K	10K	10K
Task 2		40K	10K
Task 3			50K

여러개의 다른 Task로부터 어휘적, 문맥적, 의미적 정보를 학습하기 위해 Multi-task 학습을 고려하지만 기존 방법은 아래의 단점이 존재

1) 새로운 Task를 추가하기 위해 **전체 Task를 모두 다시 학습해야한다는 효율성의 문제**

2) 따로 순차적으로 학습하는 경우 **이전 학습내용이 소실되는 문제**

- **모든 Task에 N번의 학습 iteration을 배정**
- **새로운 Task가 들어올 때 학습해오던 Task를 동시 학습**
- **이전 Task학습 파라미터를 그대로 사용**

03. Experiments

1. 실험데이터

1 Corpus 사용된 Task 정보

Corpus \ Task	Token-Level Loss			Sentence-Level Loss			
	Knowledge Masking	Capital Prediction	Token-Document Relation	Sentence Reordering	Sentence Distance	Discourse Relation	IR Relevance
Encyclopedia	✓	✓	✓	✓	✓	×	×
BookCorpus	✓	✓	✓	✓	✓	×	×
News	✓	✓	✓	✓	✓	×	×
Dialog	✓	✓	✓	✓	✓	×	×
IR Relevance Data	×	×	×	×	×	×	✓
Discourse Relation Data	×	×	×	×	×	✓	×

Table 1: The Relationship between pre-training task and pre-training dataset. We use different pre-training dataset to construct different tasks. A type of pre-trained dataset can correspond to multiple pre-training tasks.

Corpus Type	English(#tokens)	Chinese(#tokens)
Encyclopedia	2021M	7378M
BookCorpus	805M	-
News	-	1478M
Dialog	4908M	522M
IR Relevance Data	-	4500M
Discourse Relation Data	171M	1110M

Table 2: The size of pre-training datasets.

- 일부 영어데이터는 Wikipedia, Book-Corpus에서 크롤링하고 Reddit과 Discovery data를 Discourse Relation(담화관계) 데이터로 사용
- 중국어의 경우 검색엔진으로부터 뉴스, 정보검색 데이터를 수집
- 정보검색, 담화관계를 예측하는 Task를 제외하고 모든 Copus가 다양한 Task 학습에 사용됨

2. Experiments & Results on English NLP tasks

1 GLEU Benchmark Result

Task(Metrics)	*BASE model*		*LARGE model*				
	Test		Dev			Test	
	BERT	ERNIE 2.0	BERT	XLNet	ERNIE 2.0	BERT	ERNIE 2.0
CoLA (Matthew Corr.)	52.1	**55.2**	60.6	63.6	**65.4**	60.5	**63.5**
SST-2 (Accuracy)	93.5	**95.0**	93.2	95.6	**96.0**	94.9	**95.6**
MRPC (Accurary/F1)	84.8/88.9	**86.1/89.9**	88.0/-	89.2/-	**89.7/-**	85.4/89.3	**87.4/90.2**
STS-B (Pearson Corr./Spearman Corr.)	87.1/85.8	**87.6/86.5**	90.0/-	91.8/-	**92.3/-**	87.6/86.5	**91.2/90.6**
QQP (Accuracy/F1)	89.2/71.2	**89.8/73.2**	91.3/-	91.8/-	**92.5/-**	89.3/72.1	**90.1/73.8**
MNLI-m/mm (Accuracy)	84.6/83.4	**86.1/85.5**	86.6/-	**89.8/-**	89.1/-	86.7/85.9	**88.7/88.8**
QNLI (Accuracy)	90.5	**92.9**	92.3	93.9	**94.3**	92.7	**94.6**
RTE (Accuracy)	66.4	**74.8**	70.4	83.8	**85.2**	70.1	**80.2**
WNLI (Accuracy)	**65.1**	**65.1**	-	-	-	65.1	**67.8**
AX(Matthew Corr.)	34.2	**37.4**	-	-	-	39.6	**48.0**
Score	78.3	**80.6**	-	-	-	80.5	**83.6**

Table 5: The results on GLUE benchmark, where the results on dev set are the median of five runs and the results on test set are scored by the GLUE evaluation server (https://gluebenchmark.com/leaderboard). The state-of-the-art results are in bold. All of the fine-tuned models of AX is trained by the data of MNLI.

- BERT와 동일 사이즈의 모델을 활용해 base와 large 평가함. XLNET은 dev set만 공개되어 있어 dev에서만 평가
- **ERNIE 2.0 base 모델이 BERT base 보다 모든 task에서 좋은 성능 보임**
- **Large의 경우 MNLI 제외 모든 부분에서 좋은 결과를 보임**

2. Experiments & Results on Chinese NLP tasks

1 Build 8 Task Benchmark

Task	Metrics	BERT_{BASE}		ERNIE 1.0_{BASE}		ERNIE 2.0_{BASE}		ERNIE 2.0_{LARGE}	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test
CMRC 2018	EM/F1	66.3/85.9	-	65.1/85.1	-	69.1/88.6	-	**71.5/89.9**	-
DRCD	EM/F1	85.7/91.6	84.9/90.9	84.6/90.9	84.0/90.5	88.5/93.8	88.0/93.4	**89.7/94.7**	**89.0/94.2**
DuReader	EM/F1	59.5/73.1	-	57.9/72.1	-	61.3/74.9	-	**64.2/77.3**	-
MSRA-NER	F1	94.0	92.6	95.0	93.8	95.2	93.8	**96.3**	**95.0**
XNLI	Accuracy	78.1	77.2	79.9	78.4	81.2	79.7	**82.6**	**81.0**
ChnSentiCorp	Accuracy	94.6	94.3	95.2	95.4	95.7	95.5	**96.1**	**95.8**
LCQMC	Accuracy	88.8	87.0	89.7	87.4	**90.9**	**87.9**	**90.9**	**87.9**
BQ Corpus	Accuracy	85.9	84.8	86.1	84.8	86.4	85.0	**86.5**	**85.2**
NLPCC-DBQA	MRR/F1	94.7/80.7	94.6/80.8	95.0/82.3	95.1/82.7	95.7/84.7	95.7/85.3	**95.9/85.3**	**95.8/85.8**

Table 6: The results of 9 common Chinese NLP tasks. ERNIE 1.0 indicates model released by (Sun et al. 2019, ERNIE) . The reported results are the average of five experimental results, and the state-of-the-art results are in bold.

- CMRC 2018 / DRCD / DuReader: Machine Reading Comprehension (MRC)
- MSRA-NER: Named Entity Recognition (NER)
- XNLI: Natural Language Inference (NLI)
- ChnSentiCorp: Sentiment Analysis (SA)
- LCQMC / BQ Corpus: Semantic Similarity (SS)
- NLPCC-DBQA: Question Answering (QA)

2. Experiments & Results on Chinese NLP tasks

2 Benchmark Test Result

Task	Metrics	BERT_{BASE}		ERNIE 1.0_{BASE}		ERNIE 2.0_{BASE}		ERNIE 2.0_{LARGE}	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test
CMRC 2018	EM/F1	66.3/85.9	-	65.1/85.1	-	69.1/88.6	-	**71.5/89.9**	-
DRCD	EM/F1	85.7/91.6	84.9/90.9	84.6/90.9	84.0/90.5	88.5/93.8	88.0/93.4	**89.7/94.7**	**89.0/94.2**
DuReader	EM/F1	59.5/73.1	-	57.9/72.1	-	61.3/74.9	-	**64.2/77.3**	-
MSRA-NER	F1	94.0	92.6	95.0	93.8	95.2	93.8	**96.3**	**95.0**
XNLI	Accuracy	78.1	77.2	79.9	78.4	81.2	79.7	**82.6**	**81.0**
ChnSentiCorp	Accuracy	94.6	94.3	95.2	95.4	95.7	95.5	**96.1**	**95.8**
LCQMC	Accuracy	88.8	87.0	89.7	87.4	**90.9**	**87.9**	**90.9**	**87.9**
BQ Corpus	Accuracy	85.9	84.8	86.1	84.8	86.4	85.0	**86.5**	**85.2**
NLPCC-DBQA	MRR/F1	94.7/80.7	94.6/80.8	95.0/82.3	95.1/82.7	95.7/84.7	95.7/85.3	**95.9/85.3**	**95.8/85.8**

Table 6: The results of 9 common Chinese NLP tasks. ERNIE 1.0 indicates model released by (Sun et al. 2019, ERNIE) . The reported results are the average of five experimental results, and the state-of-the-art results are in bold.

- ERNIE 1.0이 BERT base 5개정도 상회하지만 학습한 정보에 비해 부족. 이는 BERT는 512 maxlen이지만 128 이기때문
- ERNIE 2.0은 BERT base보다 모든 task에서 좋은 성능을 보임

2. Expect of Sequential Multi-task learning

1 Sequential Multi-task 효과 분석

Pre-training method	Pre-training task	Training iterations (steps)				Fine-tuning result		
		Stage 1	Stage 2	Stage 3	Stage 4	MNLI	SST-2	MRPC
Continual Learning	Knowledge Masking	50k	-	-	-	77.3	86.4	82.5
	Capital Prediction	-	50k	-	-			
	Token-Document Relation	-	-	50k	-			
	Sentence Reordering	-	-	-	50k			
Multi-task Learning	Knowledge Masking	50k				78.7	87.5	83.0
	Capital Prediction	50k						
	Token-Document Relation	50k						
	Sentence Reordering	50k						
continual Multi-task Learning	Knowledge Masking	20k	10k	10k	10k	**79.0**	**87.8**	**84.0**
	Capital Prediction	-	30k	10k	10k			
	Token-Document Relation	-	-	40k	10k			
	Sentence Reordering	-	-	-	50k			

ERNIE 2.0 Framework에서 제시한 multi-task 방법의 효과를 분석하기 위해 같은 Task들로 학습 방법만 변경해 테스트 진행
모든 방법에서 학습 반복 횟수는 동일하게 설정 (모든 Task는 50K Step 학습진행)

- Continual Learning은 학습 Task들을 순차적으로 학습
- Multi-task Learning은 여러 Task들을 동시에 학습
- **ERNIE 2.0의 continual Multi-task learning은 순차적으로 학습시키되 학습 iteration을 나눠 다른 task학습 시 동시 학습**

결과적으로 ERNIE 2.0의 방법의 성능이 가장 좋은 것을 볼 수 있음

INDEX

01. Introduction

1. Ernie 3.0에 대한 소개

1 현재 SuperGlue SOTA 모델

- Human Super Glue 89.8 점을 넘은 모델이 Deberta, T5+ Meena에 이어 3번째 모델
- 10 billion 학습 parameter들을 plain texts와 large scale knowledge corpus로 구성된 4TB dataset으로 실험

2 Ernie3.0 모델의 특징

- Ernie 3.0은 입력 데이터에 대해 Universal Representation Module 이후 Task Specific Representation Module로 구성된 Continual Multi-Paradigms Unified Pre-training Framework을 제시
- Task Specific Representation Module 은 Language Understanding representation과 Language Generation representation 두 모듈로 구성됨

3 성능 향상

- NLU, NLG에서 평균 5점 정도 향상된 성능
- NER, RE, IE, Retrieval TASK의 general한 dataset에서 SOTA
- SuperGLUE, XNLI, WSC, WebText, CoPA, MultiRC, PAWS-X 등 총 57종 TASK로 성능 검증

4 모델 및 코드 공개

- Official 코드는 현재 공개되지 않음

02. Related Work

1. Knowledge Injected PLM들 (Knowledge Enhanced Models)

Knowledge Injected PLM들	Pre-trained language model에서 world knowledge(entity, relation embedding)를 결합한 형태의 모델들

1 전형적인 world knowledge로 knowledge graph가 활용되는 경우

WKLM (2019)	원 문서의 entity 에 대한 mention을 같은 entity type의 다른 entity 이름으로 대체하고, 임의로 선택된 mention 중에서 정확한 entity mention을 구별하는 것을 모델 학습 시킴.
KEPLER (2021)	Knowledge embedding과 masked language model의 결합 방법으로 Knowledge World와 masked language model representation을 같은 semantic 공간에서 정렬하여 최적화 시킴
CoLAKE (2020)	language context와 knowledge context를 word-knowledge graph에서 통합하여 Language 와 knowledge에 대한 contextualized representation을 함께 학습 함.

1. Knowledge Injected PLM들 (Knowledge Enhanced Models)

Knowledge Injected PLM들	Pre-trained language model에서 world knowledge(entity, relation embedding)를 결합한 형태의 모델들

2 world knowledge로 large-scale data에 대한 추가 annotation을 하는 방법을 사용한 경우

Ernie1.0 (2019)	Token masking에 더하여 phrase masking과 named entity masking 전략을 소개함 전체 masked phrases와 named entities를 예측하는 것은 local contexts와 global contexts간의 의존관계 해석에 도움을 줌.
CALM (2020)	개념 순서가 부정확한 문장을 찾고 수정하는 task와 신뢰할 만한 문장과 취약한 문장을 구분 짓는 task로 표현된 두 종류의 self supervised pre-training task를 통해 model을 학습 시킴
K-Adapter (2020)	Knowledge의 소스를 구분하는 학습을 위해 추가 annotation을 하여 서로 다른 knowledge 소스로 학습된 adapter를 활용함.

03. ERNIE3.0

1. Ernie3.0 Framework

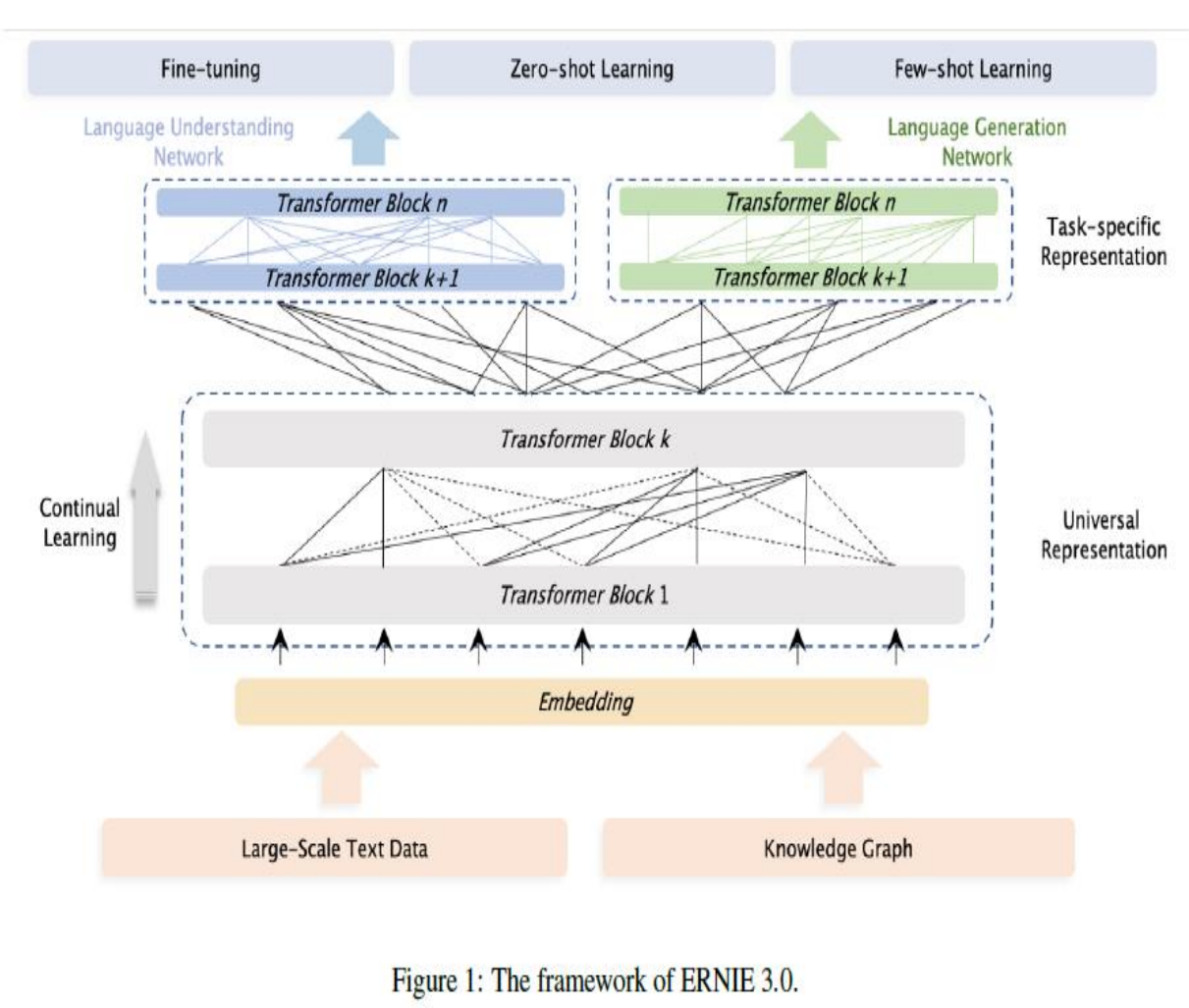


Figure 1: The framework of ERNIE 3.0.

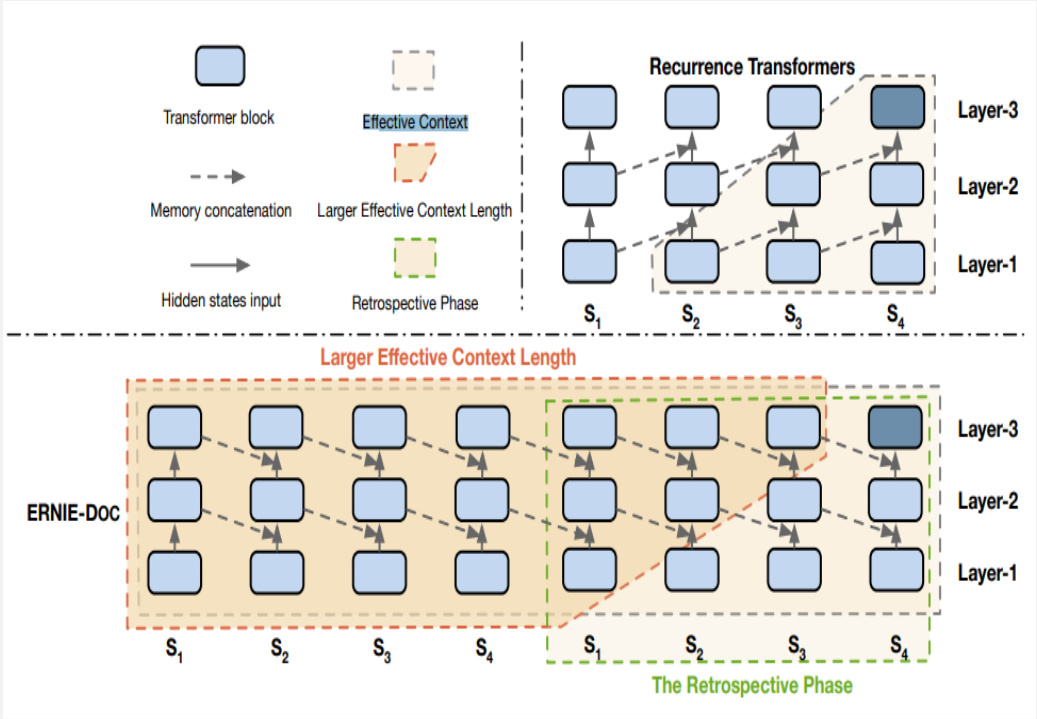
1 Continual Multi-Paradigms Unified Pre-training Framework

- **Universal Representation Module** :
  Continual Learning을 하되, long-text의 추적 가능한 **Transformer-XL** 을 backbone으로 사용함.
  long-text 추적을 위해 보조 **recurrence memory module**을 사용하는데, 이 recurrence memory module을 모든 task에 대해 공유하면서 학습하는 구조임. Universal Representation module을 통해 일반적인 Transformer 와 같이 self-attention연산을 통해 sequence에서 각 token에 대한 contextual 정보를 capture하고 **contextual embedding sequence**를 생성함.

- **Task Specific Representation Module** :
  NLU-Specific representation과 NLG-specific representation 두 모듈로 구성됨. **NLU rep. 은 bi-directional modeling network**으로 구성하고, **NLG rep.은 unidirectional modeling network**으로 구성됨.

[별첨] Transformer-XL

Transformer 모델의 제약점	정해진 길이 이상의 long-term dependency를 학습하기 어려움
Trasnformer-XL	• RNN 계열의 recurrence 방식을 사용해서 이전 segment의 hidden state를 가져와서 현재 segment의 hidden state를 생성하는 과정에 사용 • Relative positional encoding을 사용
Ernie-DOC	• ERNIE-Doc에서 제시한 memory recurrence mechanism은 Transformer-XL에서 제시한 recurrence 방식(현재 layer가 이전 layer의 hidden state를 참고하던 것)을 이전 layer가 현재 layer의 hidden state를 참고하도록 변경 "



2. Ernie3.0의 Pre-training Task

- Ernie3.0의 pre-training task는
  (1) word-aware pre-training task, **(2) structure-aware pre-training task**, 그리고 **(3) knowledge-aware pre-training task**로 나뉨

1 Word-aware pre-training task

- **Ernie1.0**에서 사용한 **word-level masking, phrase-level masking, entity-level masking**을 사용
- Ernie3.0에서는 **generative model**특성을 위해 **GPT-2**언어 모델을 pre-training model로 하여 효과성을 높임.
- **긴 문장 생성** 장점을 살리기 위해 Ernie-doc에서 제안한 enhanced recurrence memory mechanism을 사용
- **Enhanced recurrence memory mechanism**으로 한 계층 씩 아래 방향으로 recurrence 함으로써, 기존 recurrence transformer보다 더 크고 효과적인 context길이 모델을 구성함

2 Structure-aware pre-training task

- **Ernie2.0**에서 사용한 "**sentence reordering**"으로 paragraph를 m개의 segments로 나누고 이 **순서를 맞추는(k classification) 학습** 방법 사용
- **Sentence Distance**는 Next Sentence Prediction (NSP)를 확장한 방법.
  3종류의 classification 을 다루는데
  (1) 두 sentence가 인접하였는지,
  (2) 인접하지 않았지만 한 문서 안에 있는지,
  (3) 인접하지는 않았고 서로 다른 문서에서 왔는지를 구별하는 방법. 이를 학습에 적용.

2. Ernie3.0의 Pre-training Task

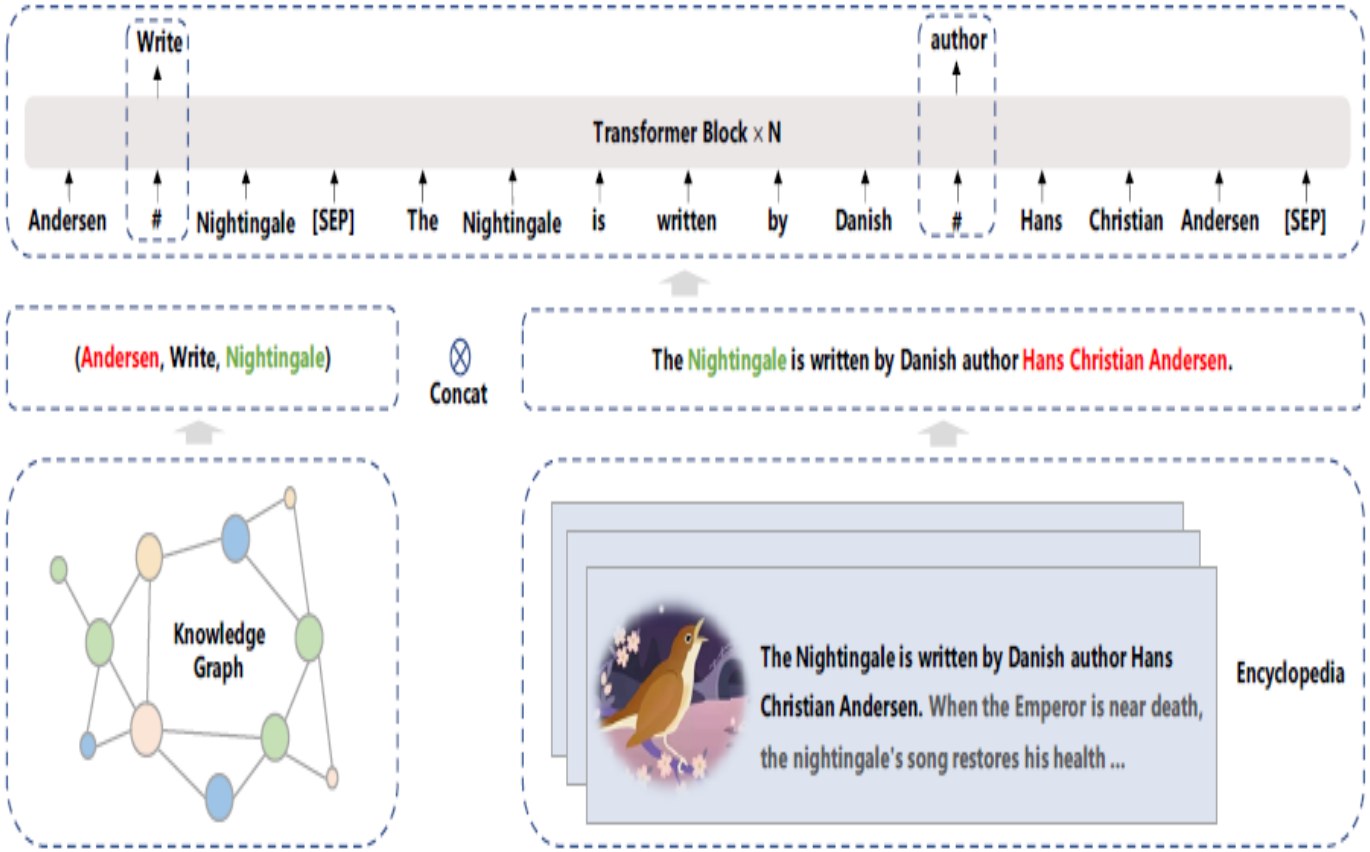


Figure 2: Universal Knowledge-Text Prediction.

2 Knowledge-aware pre-training task

- **Ernie3.0**에서는 knowledge masked language modeling을 확장하여, unstructured texts뿐만 아니라 knowledge graph도 함께 고려한 **universal knowledge-text prediction task**를 소개함
- **문장 내에 등장한 head, tail entity간의 relation을 찾는 문제**에 대해 **knowledge triple에서 head, tail에 대한 relation을 찾는 방법**을 사용함.
- 이 때 **knowledge graph에 있는 정보와 encyclopedia의 정보를 함께 학습** 시킴.
- head, tail이 한 문장에 있을 때 그 relation이 그 문장 내에 있다는 "**distant supervision task**" 가정을 따름.

3. Ernie3.0의 Pre-training Algorithm

Progressive Pre-training

- **학습의 안정성과 학습 속도 향상을 위해 효율적인 작은 모델에서 학습을 시작하다가 점차적으로 학습시키는 량을 증가시키는 방법**
- BERT는 학습 시 처음 90% 업데이트 시 줄어든 sequence 길이로 학습을 함
- [1]은 학습 시 batch 크기를 작은 값에서 큰 값으로 점점 증가 시킴
- [2]는 regularization factors(drop-out, mix-up 등)를 input 크기에 따라 stage-wise하게 변화를 주는 것이 학습 속도 향상에 도움이 된다고 확인함

ERNIE 3.0 Progressive Pre-training Algorithm

1 기존 방법들을 잘 융합한 학습 절차

- **(1) 입력 sequence의 길이,**
  (2) batch 크기,
  (3) learning rate,
  (4) drop-out rate 과 같은 regularization factor들을 점진적으로 동시에 키워가면서 학습을 진행함.
- 학습 안정성을 위해 일반적으로 사용하는 **learning warm-up** 전략도 차용함.

[1] Alec Radford et. al., "Language models are unsupervised multitask learners," Open AI Blog, 2019.
[2] Mingxing Tan et. al., "Efficientnetv2: Smaller models and faster training, CoRR, 2021.

4. Ernie3.0의 학습 데이터 및 전처리

1 학습 데이터의 구성

- Ernie2.0 구성 dataset: baike, Wikipedia, feed and etc
- Baidu Search dataset : Baijiahao, Zhidano, Tieba, Experience
- Web text, QA-long, QA-short, Poetry, Couplet,
- 의료, 법률, 경제 등의 domain specific data
- 50 million facts이상을 지닌 Baidu Knowledge graph

- Baike : formal language로 쓰인 백과사전
- Baijiahao : content creation platform의 문서들
- Zhidano: Chinese interactive Q&A community platform
- Tieba: 취미나 핫이슈에 대한 교류와 정보를 공유 하는 social platform
- Couplet : 저속한 단어를 제외시킨 중국어 커플 스토리 셋

1 Text 전처리

- **중복성 제외** : Character-level, paragraph-level, document-level로 중복되는 내용을 제외, 혹은 변형 시킴. 중복 문서 제외에 **MD5(Message Digest Algorithm5)**를 이용함.
- **10단어 이하 문장은 제외 시킴.**
- Sentence segmentation을 regular expression과 word segmentation에 맞게 수정함.
- 50 Million facts이상의 Baidu knowledge graph가 이용됨.

4. Ernie3.0의 Text 전처리 및 학습 환경

1 학습 환경

- Universal rep. module과 Task-specific rep. module을 결합한 전체 학습 parameter ~ 10 billion
- Universal Representation Module: 48 layers, 4096 hidden units, 64 heads
- Task-specific Representation Module: 12 layers, 768 hidden units, 12 heads
- Active 함수 : GeLU
- Maximum sequence Length : 512
- Maximum memory length for language generation : 128
- Total batch size : 6144
- Optimizer : Adam with learning rate 1e-4, β_1 0.9, β_2 0.999
- L2 weight decay : 0.01
- Learning rate warm-up : 10,000 steps
- Vocabulary : 375 Billion tokens
- GPU Server : 384 NVIDIA v100 GPUs
- "parameter sharing"[3],[4]을 통해 memory usage를 줄임

[3] Samyam Rajbhandari et. al., "Zero: Memory optimizations toward training trillion parameter models," In SC20, HPCNSA, IEEE, 2020.

[4] Aditya Ramesh et al., "Zero-shot text-to-image generation," 2021.

04. Experiments

1. 실험 진행에 대한 overview

Ernie3.0 모델의 성능 비교 평가를 **NLU, NLG, Zero-Shot Learning에서 총 57 개의 fine-tuning Tasks를 사용함**

27

14종 NLU Tasks by 45 dataset +3(LUGE) +8(SuperGlue) +2(UKPT)

- 비교모델들 : SKEP, Ernie2.0, Roberta, CPM-2, ZEN-2.0, BERT-GiGRU, Glyce+BERT, ALBERT, Ernie-GRAM, MacBERT (**Deberta는?**)
- 평가 metric으로 **accuracy, F1 ,F1-macro/F1-micro, EM, QAC/PAC, MRR, NDCG-1**이 이용됨

10

7종 NLG Tasks by 9 dataset +3(LUGE)

- 비교 모델들: RoBERTa-Large, ERNIE 2.0-Large, ProphetNet-코, mT5, CPM-2 (**GPT-3나 Meena는?**)
- Text Summarization, Closed Book QA는 **Rouge-L**평가 metric으로,
- Generation (Question, Advertisement, Dialogue) 은 **BLUE-4**, Translation은 **BLUE** score를 평가 metric으로 Math는 **accuracy**를 평가 metric으로 비교함.

20

7종 Zero-Shot Learning Tasks by 19 dataset +13

- 비교 모델들: CPM-1(2.6B), PANGU-α (2.6B), PANGU-α (13B), ERNIE 3.0(10B)
- 3종 **coherence, fluency, accuracy 평가 metric** 으로 비교함.

1. Experiments on Natural Language Understanding Tasks

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

1 실험에 사용된 dataset 특징

- Task별 비교 데이터 셋 중 중국어 및 다국어 dataset은 43 종 (XNLI도 영어를 중국어로 번역한 것을 사용), 영어 및 multi-lingual dataset은 semeval2016, OntoNotes 2종
- 대부분 중국어 data를 기반으로 학습하여서 중국어 부분에서 향상된 성능을 보이고 있음.

2 TASKs

- Natural Language Inference – 주어진 전제가 다른 가설을 수반하는지를 확인
- Winograd Schema Challenge – 고유명사와 대명사간의 Coreference 관계 확인
- Semantic Similarity – sentence-level similarity를 측정

1. Experiments on Natural Language Understanding Tasks

ID	Task	Dataset	Metric		Previous SoTA Model	ERNIE 3.0
8	Chinese News Classification	TNEWS	Acc.	Dev	58.32 (RoBERTa*)	69.94
		IFLYTEK	Acc.	Dev	62.75 (RoBERTa*)	63.45
		THUCNEWS	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

2 TASKs

- Closed-book question answering – 외부 참조나 지식없이 질의에 직접 응답을 확인
- Cant understanding – 암호해독 게임 dataset인 DogWhistle을 이용하여 비교

3 Metrics

- MRR(Mean Reciprocal Rank) : 추천 등에서 제안된 순서의 역수 값을 합하여 전체 샘플개수로 나눈 평균 값으로 rank 정확도도 함께 고려한 metric.

1. Experiments on Natural Language Understanding Tasks

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**
		$\text{DuReader}_{\text{robust}}$	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**
		$\text{DuReader}_{\text{checklist}}$	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**
		$\text{DuReader}_{\text{yesno}}$	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*).

2 TASKs

- Machine Reading Comprehension(독해력) 평가에서는 띄어쓰기 예측 독해력, 다중선택지 질의 응답 독해력, 문장완성, 일관성 등에 대한 평가가 이뤄짐.
- Legal Document Analysis – 다중 labeling된 법률 문서 분류를 수행

3 Metrics

- QAC(Question-level Accuracy):

$$\frac{\#correct\ prediction}{\#total\ blanks\ in\ dataset} \times 100\%$$

- PAC(Passage-level Accuracy):

$$\frac{\#correct\ passages}{\#total\ passages\ in\ dataset} \times 100\%$$

- F1-macro는 평균의 평균, F1-micro 는 클래스별 개체의 개수를 고려한 평균, 클래스별 개체의 개수가 서로 다를 때는 F1-micro가 더 정확한 수치로 고려됨.
- nDCG(normalized Discounted Cumulative Gain) : 0과 1범위 값으로 추천 시스템의 순서별로 상위권의 중요도까지 고려한 normalized된 평가 metric
- EM은 각 질의 set의 정답과 정확한 match를 평가

2. Experiments on Natural Language Generation Tasks

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**
	$\text{DuReader}_{\text{robust}}$-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**

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

1 TASKs

- Question Generation은 기계 독해의 역 task로 주어진 답변에 대해 합리적인 질문을 만들어 내는 task
- Math는 초등학교 산술문제 풀이를 직접 수행하는 task로 23161개의 문제 묘사와 방정식과 답변으로 구성된 Math23K dataset이용

2 Metrics

- Rouge-L(Recall Oriented Understudy for Gisting Evaluation)은 Chin-Yew Lin이 고안한 방법으로 텍스트 자동요약, 기계 번역 등 자연어 생성 모델 평가를 위한 지표로, 모델의 결과와 사람의 결과를 비교함. Gold standard(사람의 결과)대비 기계의 응답을 LCS기법으로 최장길이 문자열에 대해 precision과 recall을 종합한 점수로 평가

3. Experiments on LUGE BenchMark

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	$\text{DuReader}_{\text{robust}}$	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	$\text{DuReader}_{\text{robust}}$-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.

- LUGE는 한국의 KLUE와 유사한 중국어 자연어처리 bench mark
- 6종류의 task에서 Ernie3.0은 평균 5.36% 향상된 성능을 확인

4 Experiments on Zero-shot Learning(1) - Evaluation

- gradient update나 fine-tuning 없이 모델을 적용하는 다양한 zero-shot task를 수행
- ERNIE3.0은 최근에 제안된 CPM-1, PanGu-α-2.6B, PanGu-α-13B와 같은 **large-scale LM보다 대부분의 task에서 더 좋은 성능을 보임**
- ERNIE3.0은 13개의 task에서 수집한 450개 종류의 case에서 **더욱 일관되고, 자연스럽고, 정확한 응답을 생성**할 수 있음

[Evaluation]

1 Perplexity-based Method

- CHID, CMRC2017과 같이
  여러 후보 중 한 개의 정답을 찾는 task에서
  context의 blank를 채울 때
  token당 perplexity 점수를 계산
- **더 낮은 token별 perplexity score를 정답으로 예측**
- Binary, multiple classification에서
  semantically meaningful한 label을 할당하고
  prompt를 사용해
  context와 label을 사람이 읽을 수 있는 텍스트로 형식화 함
- 이러한 task는 multi-choice task로 다루어짐

2 Generation-based Method

- Closed-book QA와 같은 자유 형식으로 완성하는 task에서
  길이 penalty없이 beam width 8인 beam search를 사용
- 완성 시 최대 생성 길이는
  데이터셋 정답 길이의 95% percentile을 기반으로 정의됨
- metric은 exact match, F1, Rouge-1이 사용됨
- extractive MRC같은 제한된 완성에서
  이전처럼 같은 파라미터로 제한된 beam search를 사용
- 각 샘플마다 구성되는 **Trie-Tree**는
  생성 공간을 효율적이고 효과적으로 적제하고,
  주어진 텍스트 내에서 발생하는 completion만을 생성함

4 Experiments on Zero-shot Learning

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.

4 Experiments on Zero-shot Learning(2) – Results(1)

1 Chinese News Classification

- 샘플 당 임의로 세 개의 negative label을 뽑고, 이 세 개의 token별 perplexity score를 비교해, 각 후보의 score를 계산 비용을 줄임
- **TNEWS에서 이전에 fine-tuning 방식으로 SOTA를 달성한 모델보다 성능이 좋고,**
  IFLYTEK에 대해서도 조금 더 성능이 좋음

2 Semantic Similarity

- **AFQMC, CSL 데이터셋에서 baseline보다 훨씬 성능이 좋음**
- **정확도는 random-guess보다 조금 더 좋은 정도인데,**
  prompt의 차선책 선택 때문일 수 있음

3 Natural Language Inference

- OCNLI, CMNLI 데이터 셋을 사용
- prompt를
  $SENT_A? NO/YES/MAYBE, $SENT_B
  로 사용
- **baseline과 비교할 만한 수준이며, pre-trained model의 zero-shot NLI task에 개선할 여지가 있음**

4 Winograd Schema Challenge

- WSC2020에서 per-token perplexity를 계산하기 위해 대명사를 각 후보로 치환해
  multi-choice completion task로 변형
- PanGu-α-13B에 비해 3.38%p 더 좋은 성능을 보임

4 Experiments on Zero-shot Learning(2) – Results(2)

5 Cloze and completion

- CHID에서 각 문장이 하나의 blank만 포함하도록 해
  객관식 정답 선택 task로 정형화 함
- **ERNIE3.0은 가장 높은 점수를 달성**

- Chinese WPLC는 masked text와 정답으로 구성됨
- PanGu-α에 비해 ERNIE3.0은
  더 낮은 perplexity score를 달성

- CMRC2019에서 각 blank마다
  세 개의 negative 후보를 임의로 샘플링해
  최적의 path를 구할 수 있도록 bream search를 적용

- PD, CFT, CMRC2017도 multi-choice task로 정형화 했고
  ERNIE3.0은 **큰 차이로 baseline보다 좋은 성능을 보임**

6 Machine Reading Comprehension

- C3는 multi-choice MRC task로,
  질문, 정답 후보, 문서로 구성됨
- CMRC2018, DRCD, DuReader에서
  문서, 질문, 정답으로 구성
- CMRC2018, DRCD, DuReader에서
  큰 차이로 baseline보다 좋은 성능을 보임

7 Clozed-book Question Answering

- 모델이 pre-training에서 배운 지식으로
  정답을 생성해야 하는 task
- 입력은 MRC와 비슷하지만 문서가 없는 형태
- **WebQA에서 ERNIE 3.0은**
  baseline에 비해 좋은 성능을 보임

4 Experiments on Zero-shot Learning(3) – Case Study(1)

- 5개 타입(Question Answering, Interpretation, Dialogue, Text Generation, Summarization)의 13개 task에서 수동으로 수집한 450개의 case에서 **zero-shot generation 능력을 평가**
- human evaluation에서 annotator들은 generation quality를 [0, 1, 2]로 표시하도록 함

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])

- **Coherence** : context와 관련이 있고 일관성이 있는지 여부를 판단
- **Fluency** : 생성된 text가 읽기에 자연스러운지 여부를 판단.
  fluent text는 생성된 text 사이에서 의미적 모순이 없어야 함
- **Accuracy** : 생성된 text가 ground truth와 동일한가

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.

4. Experiments on Zero-shot Learning(3) – Case Study(2)

- zero-shot generation sample

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 Dictornary)	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.

5. Experiments on SuerGLUE

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).

- **BoolQ**: 각 예제가 짧은 구절과 예/아니오로 구성된 QA 과제로 평가 metric은 accuracy
- **CB(Commitment Bank)**: 불균형 말뭉치에 대한 자연어 추론과제로 평가 metric은 accuracy, macro-F1
- **COPA((Choice of Plausible Alternatives)** : 일반적인 상식을 기초로 인과적인 추론을 하는 과제. 데이터는 blog와 encyclopedia로 부터 주어지는데, 평가는 accuracy로 함.
- **MultiRC (Multi-Sentence Reading Comprehension)**: 다중 문장 독해는 각 예가 문맥 문단, 해당 문단에 대한 질문 및 가능한 답변 목록으로 구성된 QA 과제로 F1과 EM으로 평가.
- **ReCoRD(Reading Comprehension with Commonsense Reasoning Dataset)** : 뉴스 기사와 클로제 스타일의 질문에 대해 모델이 답변을 완료하기 위해 엔티티를 객관식으로 선택하는 과제로 F1, EM으로 평가
- **RTE(Recognizing Textual Entailment)**: 말뭉치에서 추론능력을 확인하는 과제로 accuracy로 평가
- **WiC(Word in Context)** : 문장내 단어의 정확한 의미를 이진 분류하는 것으로 accuracy로 평가
- **WSC(Winograd Schema Challenge)** : 고유명사와 대명사간의 coreference resolution task로 객관식으로 선택하는 과제로 accuracy로 평가

05. Analysis

1. Analysis (1)

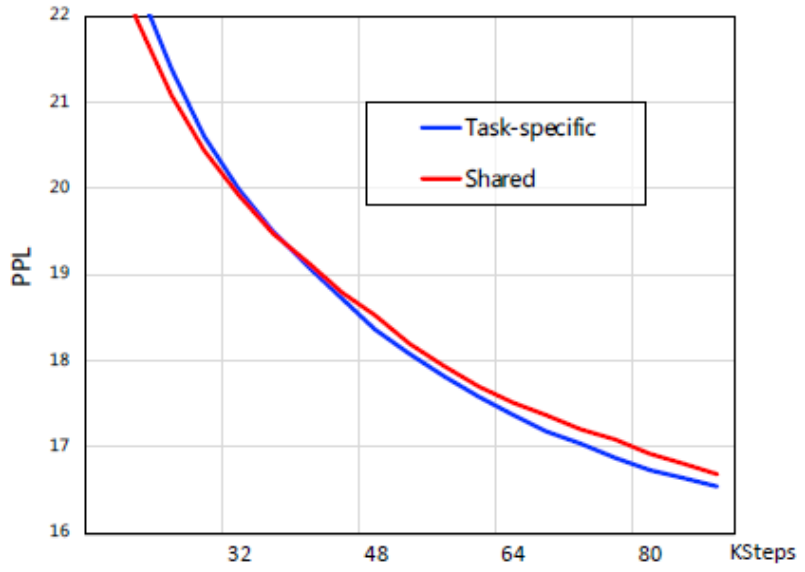


Figure 3: Perplexity variation of the NLG pre-training task with respect to training steps.

- Figure3를 통해 확인할 수 있듯이 학습을 하면서 shared parameter를 구성하는 것보다 **task-specific module을 두어서 학습을 하는 것이 더 빨리 convergence하고 더 나은 성능**에 이름

Dataset	$ERNIE_{Base}$	$ERNIE_{Base}$+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.

- **UKTP(Universal Knowledge-Text Prediction)**을 위해 Knowledge 측정에 일반적으로 사용되는 Relation Extraction Task(head, tail 로 주어진 두 entit간의 relatio을 추론)를 수행함.
  이를 위해 head와 tail의 시작과 끝을 표현한 4종류의 special token [HD],[/HD],[TL],[/TL]을 추가하여 문장내의 Head, Tail entity의 relation을 추론하는 것을 수행 하였을 때 Table10을 통해 확인 가능 하듯이 더 나은 성능을 확인
- Table5의 CKBQA task에서 Ernie3.0의 우수한 성능도 UKTP에서의 강점을 입증

1. Analysis (2)

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

Table 11: Progressive Learning To Speedup Training.

- **Progressive Learning을 적용하여 Ernie base의 경우 65.2%의 학습시간을 절감할 수 있었고, Ernie1.5B의 경우 48.2%의 학습속도 개선을 볼 수 있었음.**

- Ernie_base : 12 Layer, 768 Hidden sizes, 12 attentions heads
- Ernie 1.5B : 48 Layers, 1,536 hidden sizes, 24 attention heads

- **Progressive Learning을 적용하여 학습 속도를 개선함**
- Ernie_base와 Ernie 1.5B에 대해 8개의 V100 GPU를 이용하여 학습을 진행함.

- Ernie base에 대해
  Batch Size를 8에서 2048까지,
  Sequence 길이를 128에서 512까지,
  learning rate을 0에서 1e-4까지,
  dropout을 progressive warmup stage에서는 0으로 둠
- Ernie 1.5B에 대해
  Batch Size를 8에서 8192까지,
  learning rate을 0에서 6e-4까지
  dropout을 progressive warmup stage에서는 0으로 둠
  memory 사이즈 제한으로 gradient accumulation 전략을 사용함.

06. Conclusion

Ernie3.0은 10 Billion 학습 parameter에 대해 **Plain texts뿐만 아니라 Knowledge Graph도 함께 고려**한 학습을 함.

Pre-training시 **Task Specific Learning Module**을 두되
NLU, NLG를 모두 고려하여 학습 시킴으로써
NLU와 NLG에서 향상된 성능을 확인

Knowledge Graph도 통합한 Knowledge Enhanced 학습 전략을 통해
NLU, NLG, zero-shot learning의 각종 task에서 우수한 성능 검증

Progressive Learning학습 방법을 통해 Ernie 학습 속도 향상에 대한 기여방법도 제시됨.

감사합니다.