CTW
AMSTRAAT

ERNIE

Enhanced Representation through Knowledge Integration

양승무

CTW
AMSTRAAT

Contents

1. Introduction
2. Related Work
3. Methods
4. Experiment
5. Conclusion

Introduction

- Language representation pre-training이 NLP task에서 효과적인 방법을 보여줌
- 다양한 전략으로 여러 작업들이 이 분야를 발전시키고 있음(ELMO, BERT, GPT..)
- 이전에 나온 모델들은 문장 내에 사전 지식 정보(knowledge integration)들을 고려하지 않음
- 사전 지식 정보들이 주어지면 모델들은 더 나은 성능을 보일 것이라고 생각
- ERNIE는 knowledge masking strategies를 사용하여 BERT를 개선함
- heterogeneous Chinese data를 사용해서 다섯 가지의 NLP task에 적용함

2. Related work

2.1 Context-independent Representation

- 전통적인 방식은 문맥을 고려하지 못한 word embedding
- 한 단어가 여러 의미를 가지고 있음에도 불구하고, 한 가지 의미만 표현하는 단점이 있음
- ex) Word2Vec(Mikolov et al., 2013), Glove(Pennington et al., 2014)

2. Related work

2.2 Context-aware Representation

문맥을 고려해서 의미파악을 하는 것에 중점

- Skip-thought (Kiros et al., 2015)
- ULMFit (Howard and Ruder, 2018)
- ELMo (Peters et al., 2018)
- GPT (Radford et al., 2018)
- **BERT (Devlin et al., 2018)**
- MT-DNN (Liu et al., 2019)

3. Methods

3.1 Transformer Encoder

- GPT, BERT, XLM과 동일한 기본 encoder 사용
- 중국어 코퍼스의 경우 CJK 유니코드 범위의 모든 문자 주위에 공백을 추가하고 워드피스(WordPiece, Wu et al., 2016)를 사용하여 중국어 문장을 토큰화함
- 주어진 토큰의 경우, input representation, 해당 토큰, segment 및 position embeddings를 합쳐서 구성됨
- 모든 시퀀스의 첫 번째 토큰은 특수 분류 임베딩([CLS])으로 시작

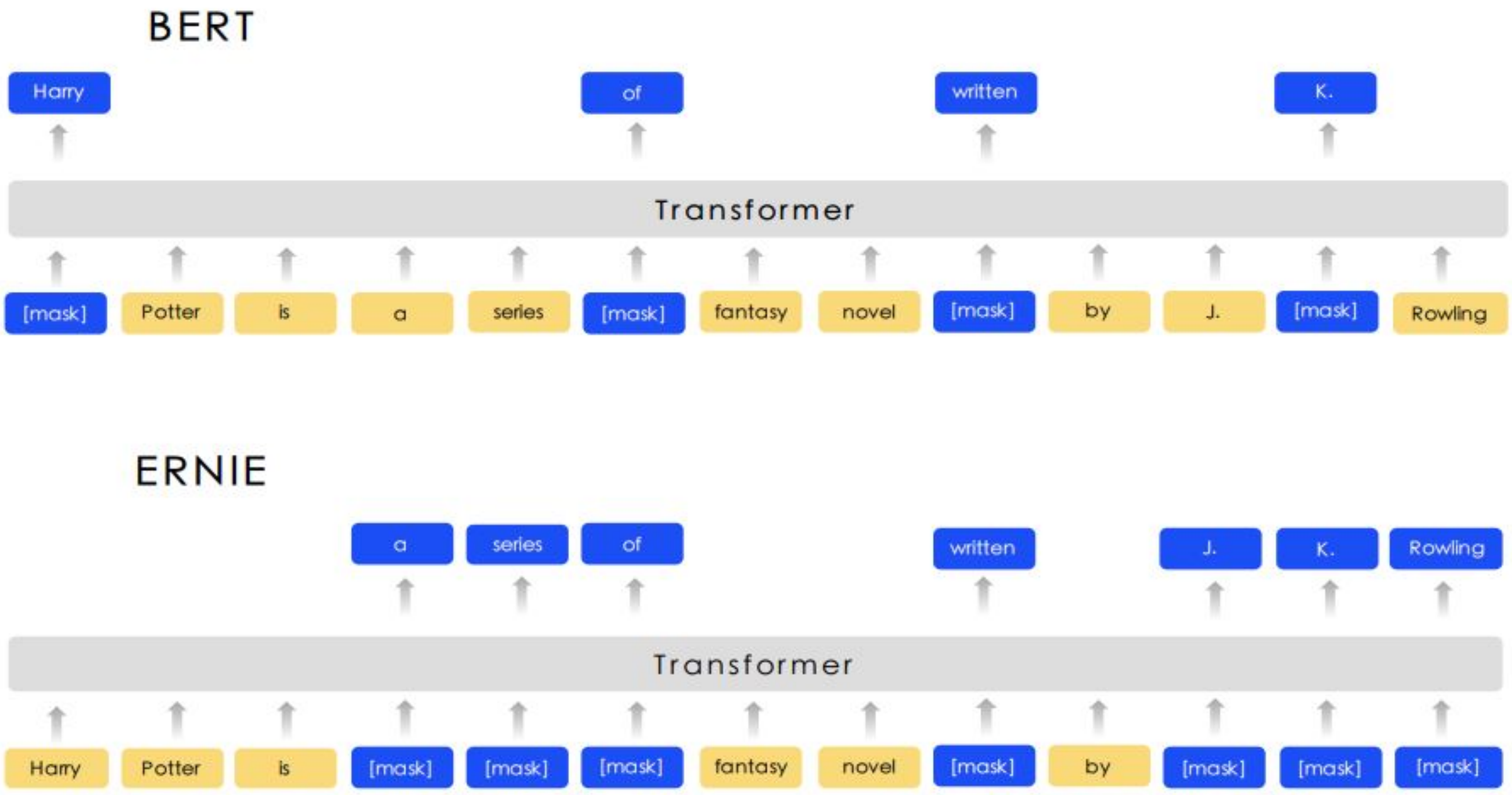


Figure 1: The different masking strategy between BERT and ERNIE

3. Methods

3.2 Knowledge Integration

- Knowledge embedding을 직접 넣지 않고 Multi-stage masking 전략을 사용

- 3.2.1 Basic-Level Masking: 기존 BERT와 동일함
- 3.2.2 Phrase-Level Masking: lexical(어휘) analysis를 통해 phrase를 추출해서 masking
- 3.2.3 Entity-Level Masking: NER을 통해 entity 추출해서 masking

Basic-Level Masking

- 기존 버트와 동일함 15 % random masking

- 기본 의미 단위의 랜덤 마스크로 훈련하기 때문에 높은 수준의 의미는 파악하기 어려움

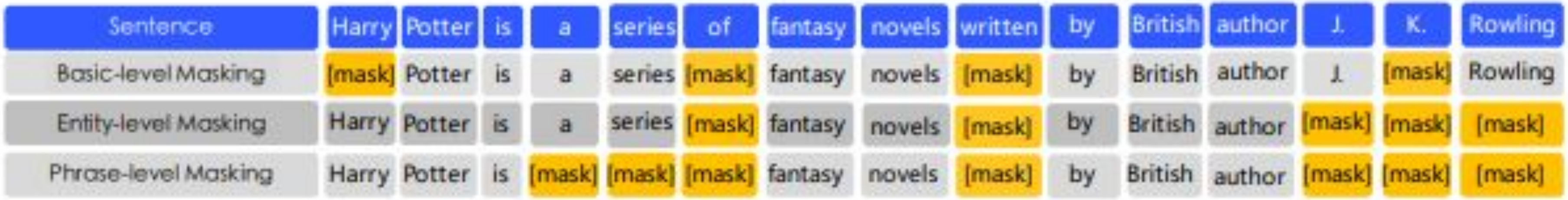
Phrase-Level Masking

- Phrase(구) 단위로 masking
- lexical analysis와 chunking tool을 사용해서 구의 경계를 구분
- segmentation tool을 이용해서 다른 언어로 단어/문자의 정보를 얻음

Entity-Level Masking

- Name entities는 인물, 장소, 기관, 제품 등 문맥에 중요한 단서가 되는 단어들을 포함
- 보통 entities는 문장에서 중요한 정보들을 포함하고 있음
- phrase masking stage에서 먼저 named entities를 분석하고, masking하여 entities 안에 있는 모든 slots를 예측함

세 단계의 Masking을 실시하고 학습하면 더 풍부한 의미 정보에 의해 강화된 단어 표현을 얻게됨



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]

Figure 2: Different masking level of a sentence

4. Experiments

4.1 Heterogeneous Corpus Pre-training

mixed corpus Chinese data

The number of sentences are 21M, 47M, 54M.

Baidu Baike: encyclopedia articles, a strong basis for language modeling

Baidu news: latest information about movie names, actor names, football team names, etc.

Baidu Tieba: open discussion forum, can be regarded as a dialogue thread.

4. Experiments

4.2 DLM (Dialogue Language Model)

Dialogue data는 semantic representation 에서 중요한 역할을 함

ERNIE의 Dialogue 임베딩은 ERNIE가 multi-turn conversation을 가능하게 함

DLM task가 ERNIE로 하여금 함축적인 관계를 파악하는 데 도움을 줌

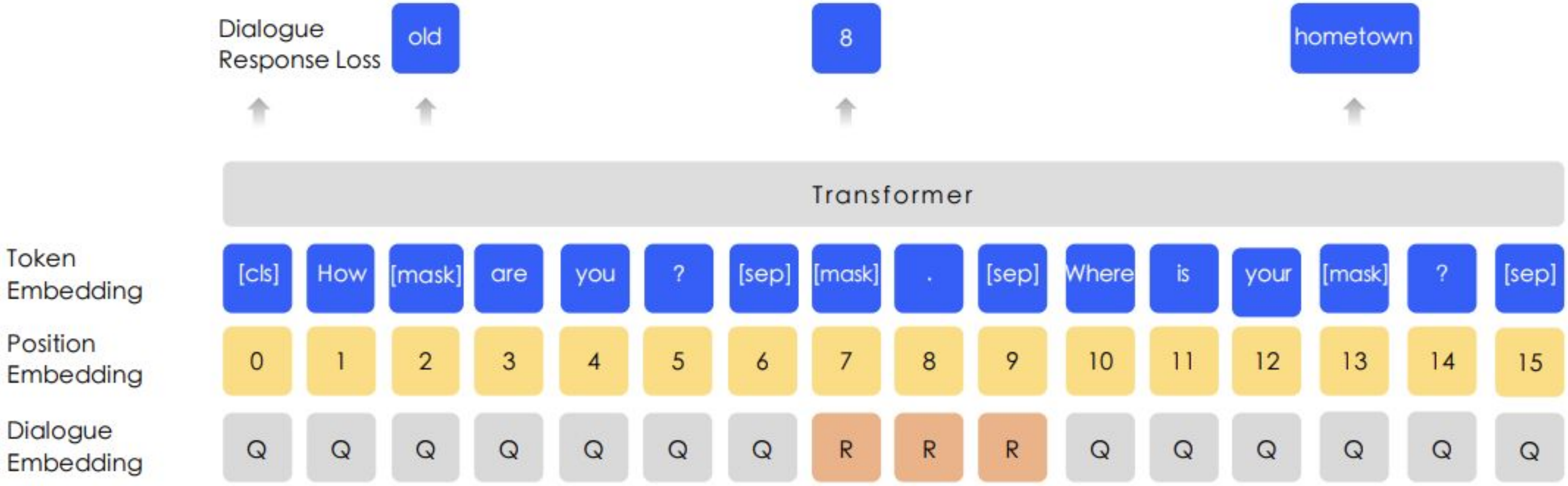


Figure 3: Dialogue Language Model. Source sentence: [cls] How [mask] are you [sep] 8 . [sep] Where is your [mask] ? [sep]. Target sentence (words the predict): old, 8, hometown)

4.3 Experiments on Chinese NLP Tasks

4.3.1 Natural Language Inference (XNLI)

4.3.2 Semantic Similarity (LCQMC)

4.3.3 Named Entity Recognition (MSRA-NER)

4.3.4 Sentiment Analysis (ChnSentiCorp)

4.3.5 Retrieval Question Answering (NLPCC-DBQA)

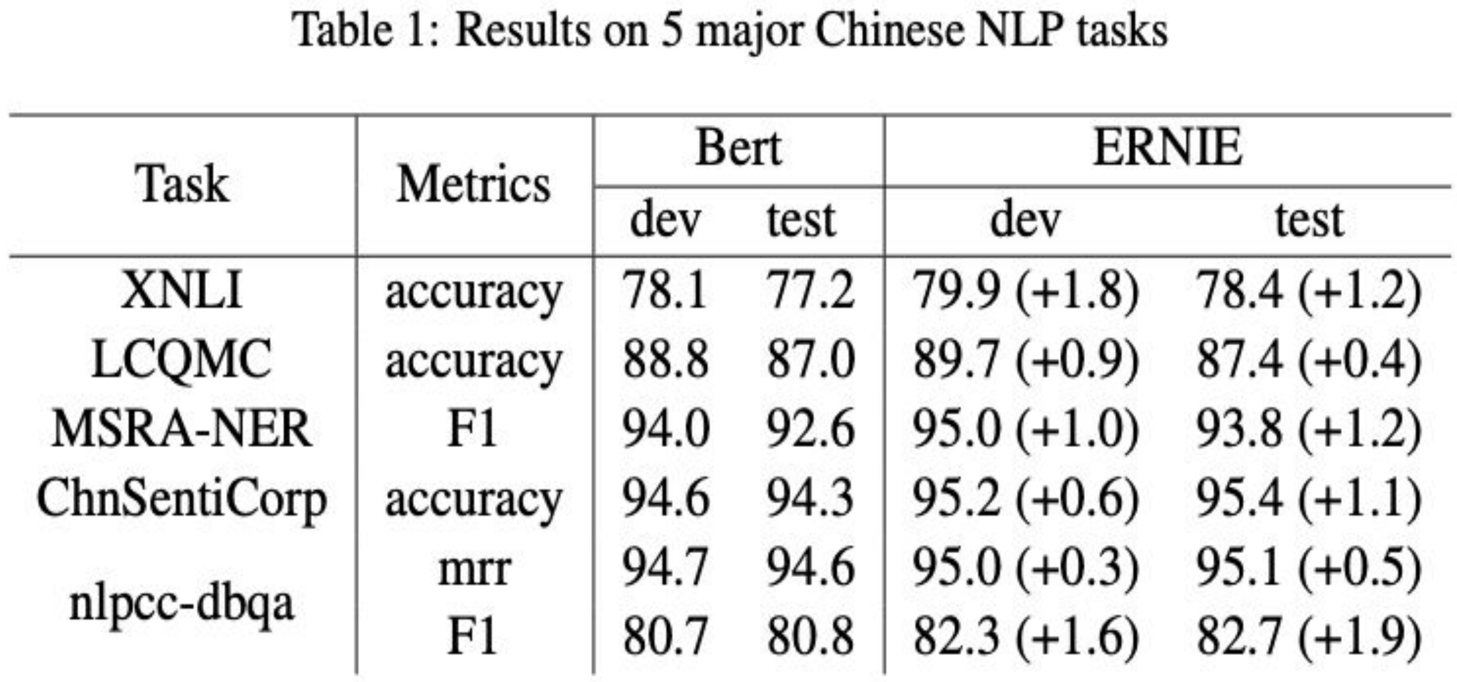


Table 1: Results on 5 major Chinese NLP tasks

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)

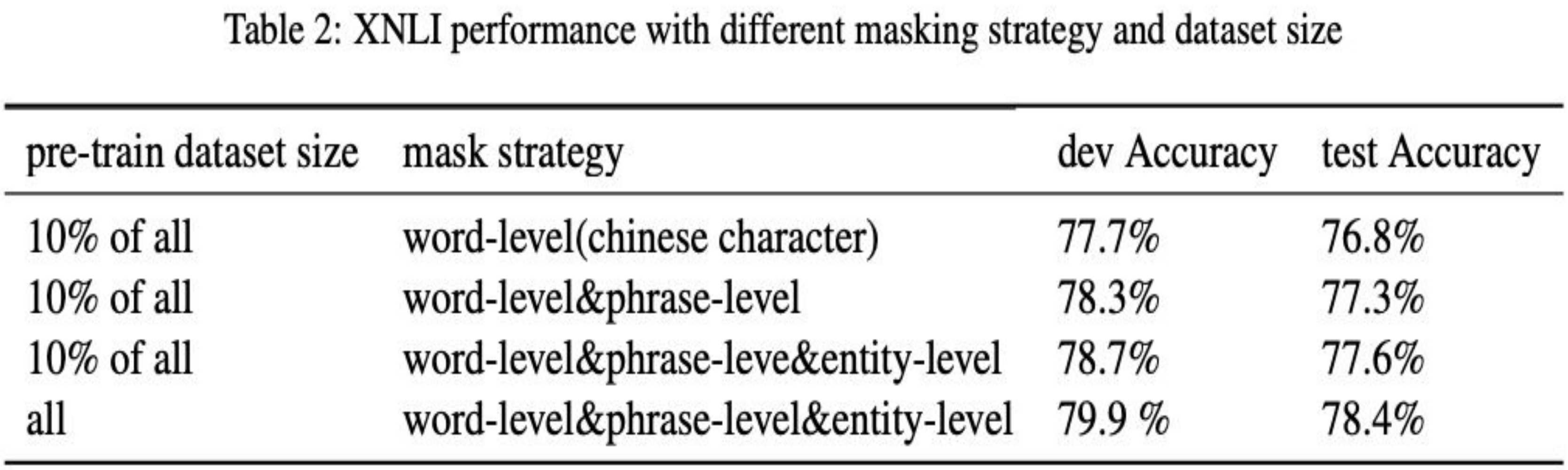


Table 2: XNLI performance with different masking strategy and dataset size

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%

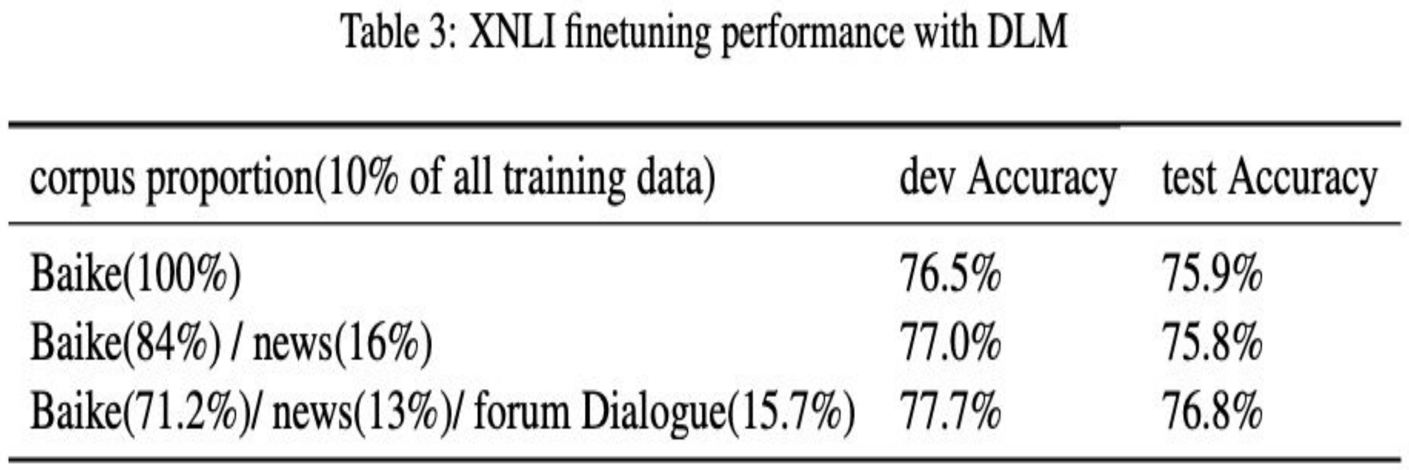
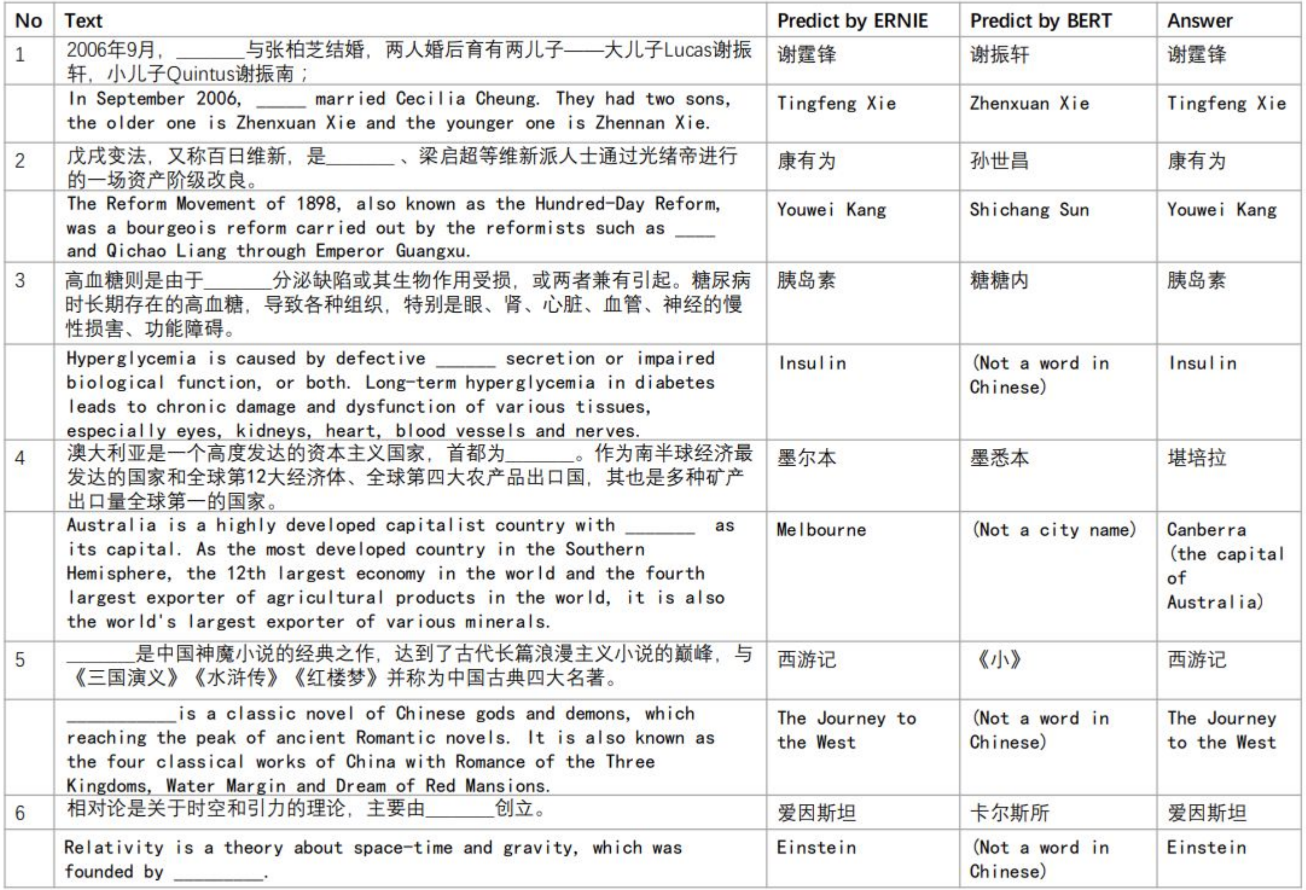


Table 3: XNLI finetuning performance with DLM

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%



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

Conclusion

- 사전 지식 정보들을 pre-training language model에 적용하여 모델의 향상을 이끌어냄
- 자신들의 방법이 모든 task에서 BERT에 비해 뛰어나다고 함
- knowledge integration 방법과 heterogeneous data에 pre-training을 적용하면 모델이 더 나은 성능을 보인다고 함