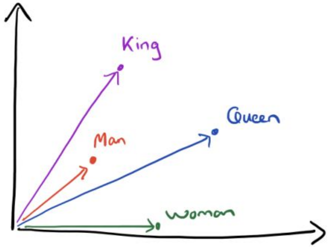
ERNIE: Enhanced Representation through Knowledge Integration

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, Hua Wu

Baidu Inc. (2019)

Pre-trained language representations

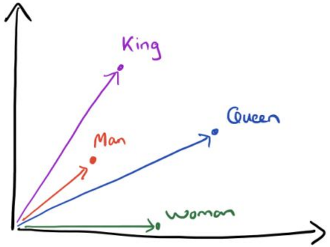


Word2Vec

ELMo

BERT

Pre-trained language representations



Word2Vec

ELMo

BERT

ERNIE

BERT

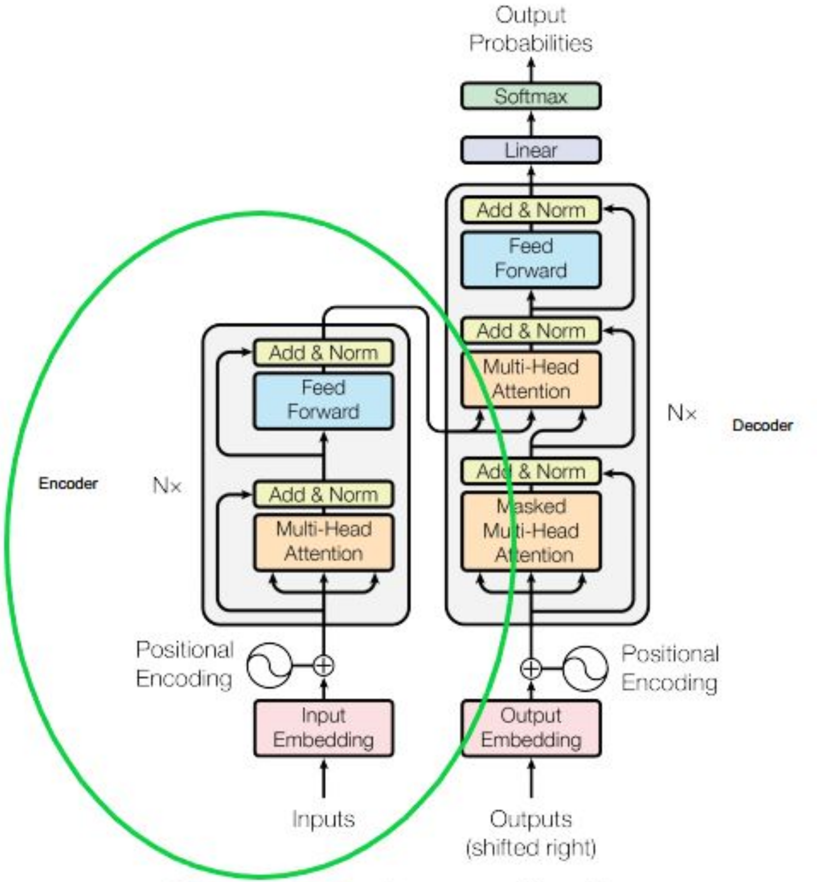
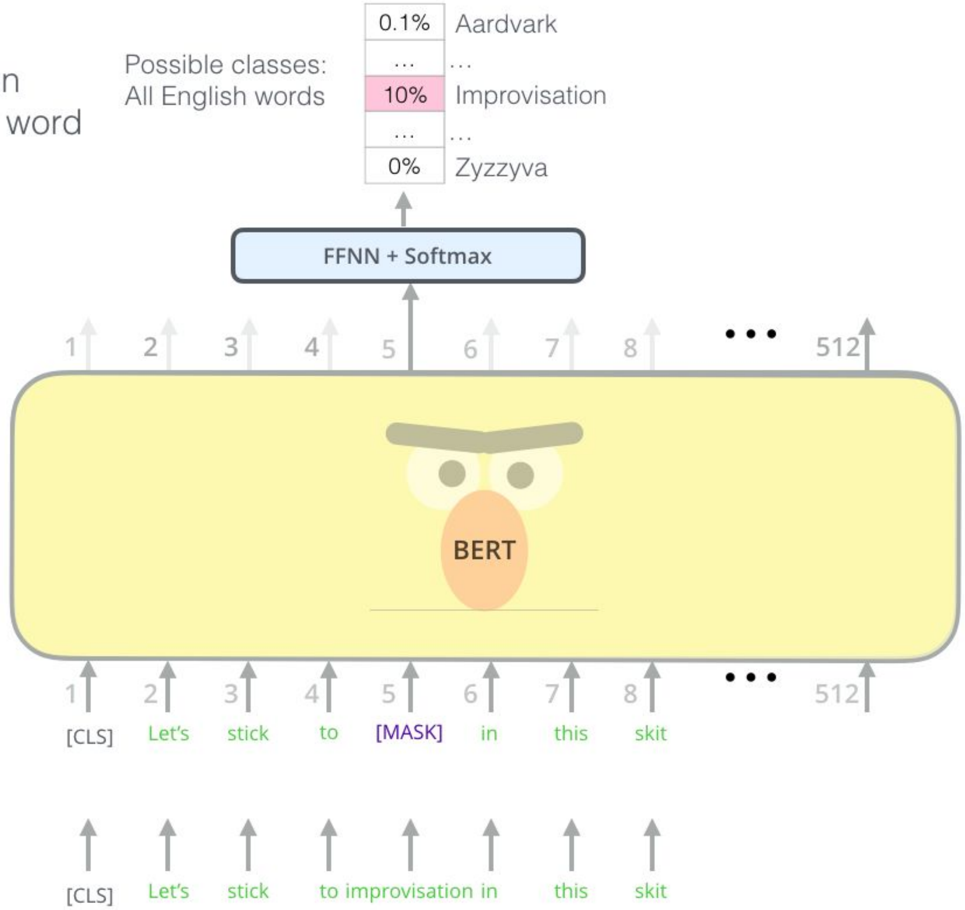


Figure 1: The Transformer - model architecture.

BERT

Use the output of the masked word's position to predict the masked word



Possible classes: All English words

0.1%	Aardvark
…	…
10%	Improvisation
…	…
0%	Zyzzyva

Randomly mask 15% of tokens

Input

Same As BERT but ...

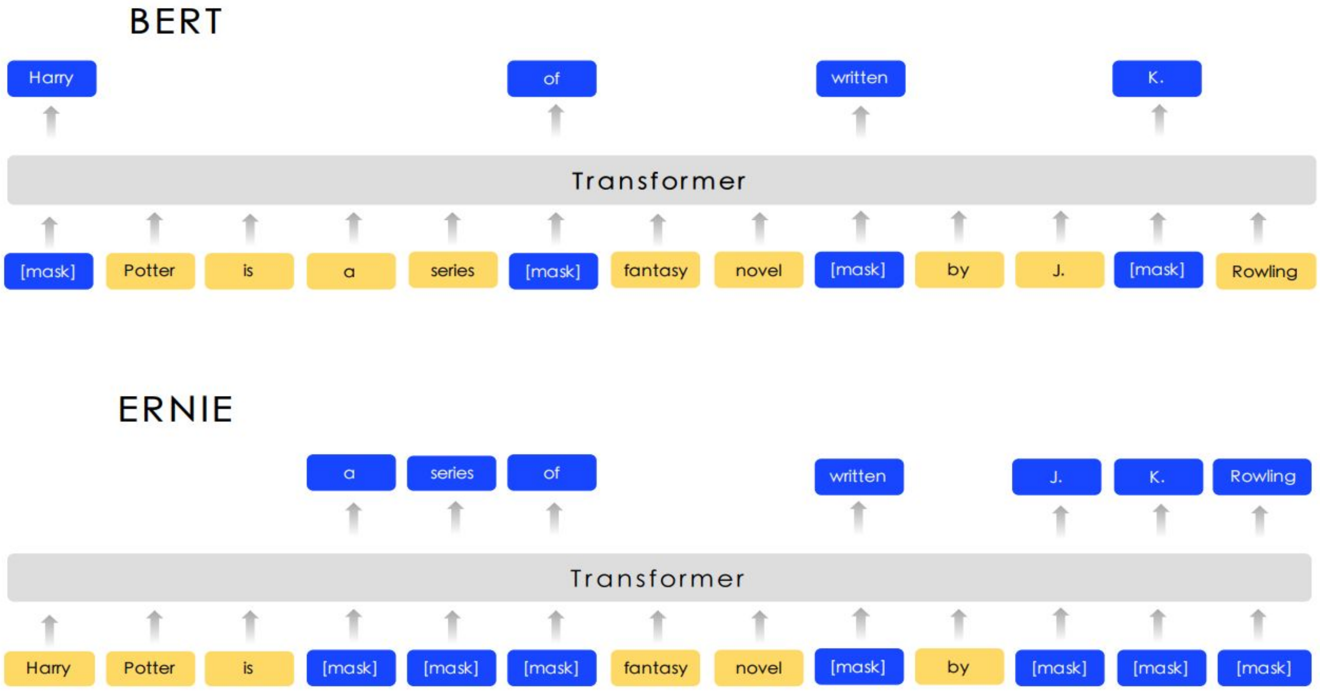
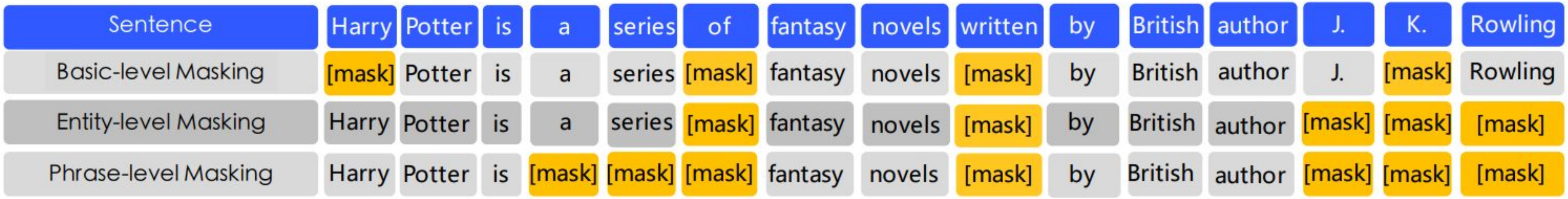


Figure 1: The different masking strategy between BERT and ERNIE

Different masking strategies

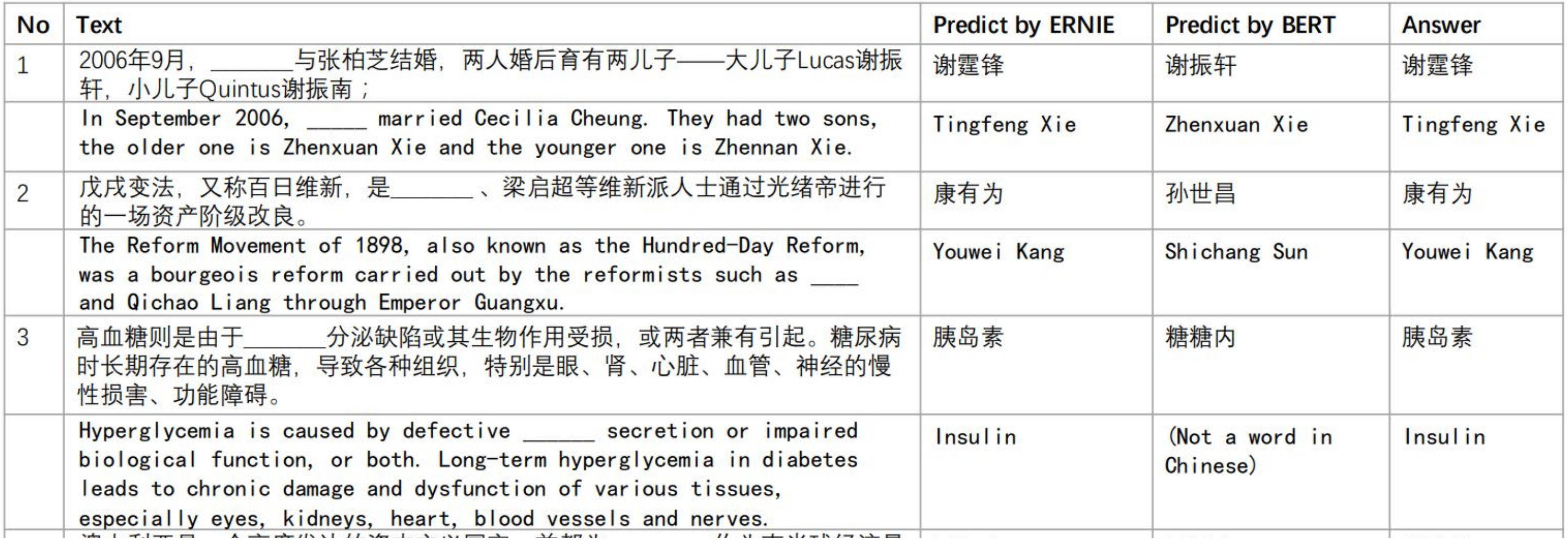
We use prior knowledge to enhance our pretrained language model. Instead of adding the knowledge embedding directly, we proposed a multi-stage knowledge masking strategy to integrate phrase and entity level knowledge into the Language representation.



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

Examples



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

Results

ERNIE was chosen to have the same model size as BERT-base for comparison purposes. ERNIE uses 12 encoder layers, 768 hidden units and 12 attention heads.

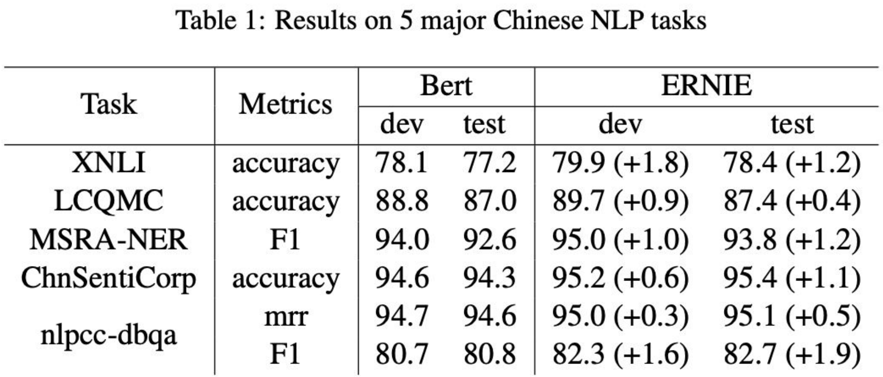


Table 1: Results on 5 major Chinese NLP tasks

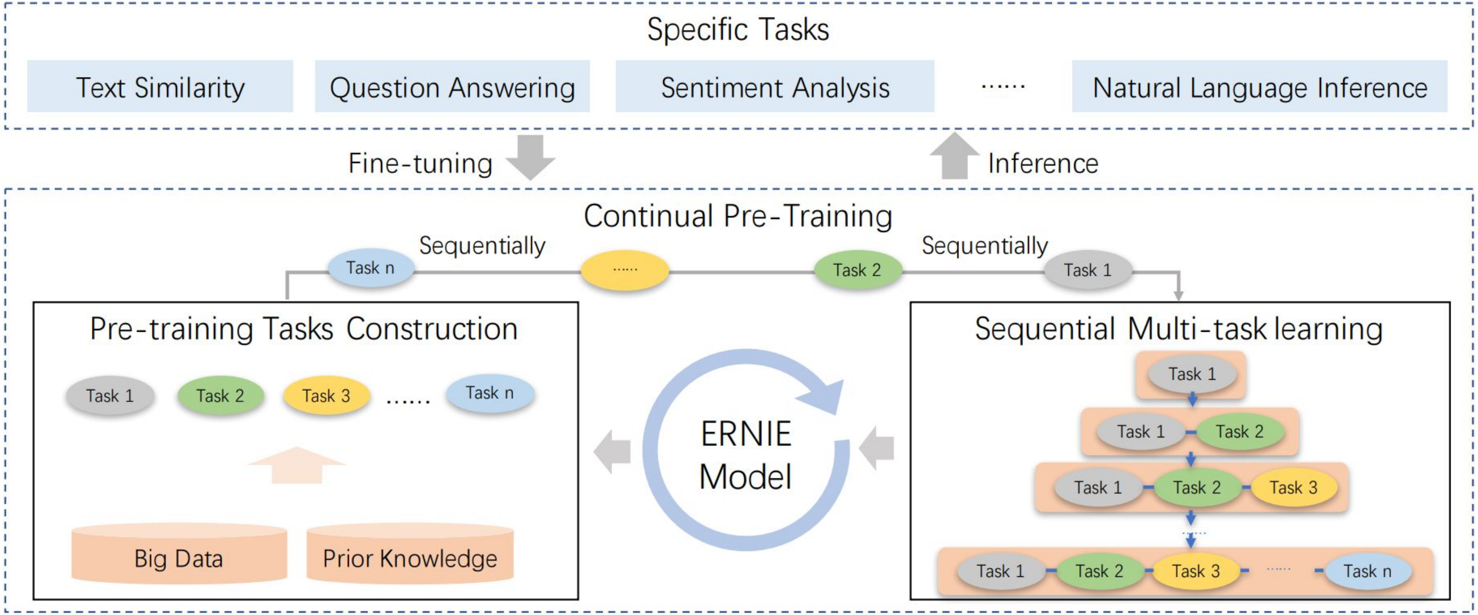
Task	Metrics	Bert		ERNIE	
		dev	test	dev	test
XNLI	accuracy	78.1	77.2	79.9 (+1.8)	78.4 (+1.2)
LCQMC	accuracy	88.8	87.0	89.7 (+0.9)	87.4 (+0.4)
MSRA-NER	F1	94.0	92.6	95.0 (+1.0)	93.8 (+1.2)
ChnSentiCorp	accuracy	94.6	94.3	95.2 (+0.6)	95.4 (+1.1)
nlpcc-dbqa	mrr	94.7	94.6	95.0 (+0.3)	95.1 (+0.5)
	F1	80.7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

ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding

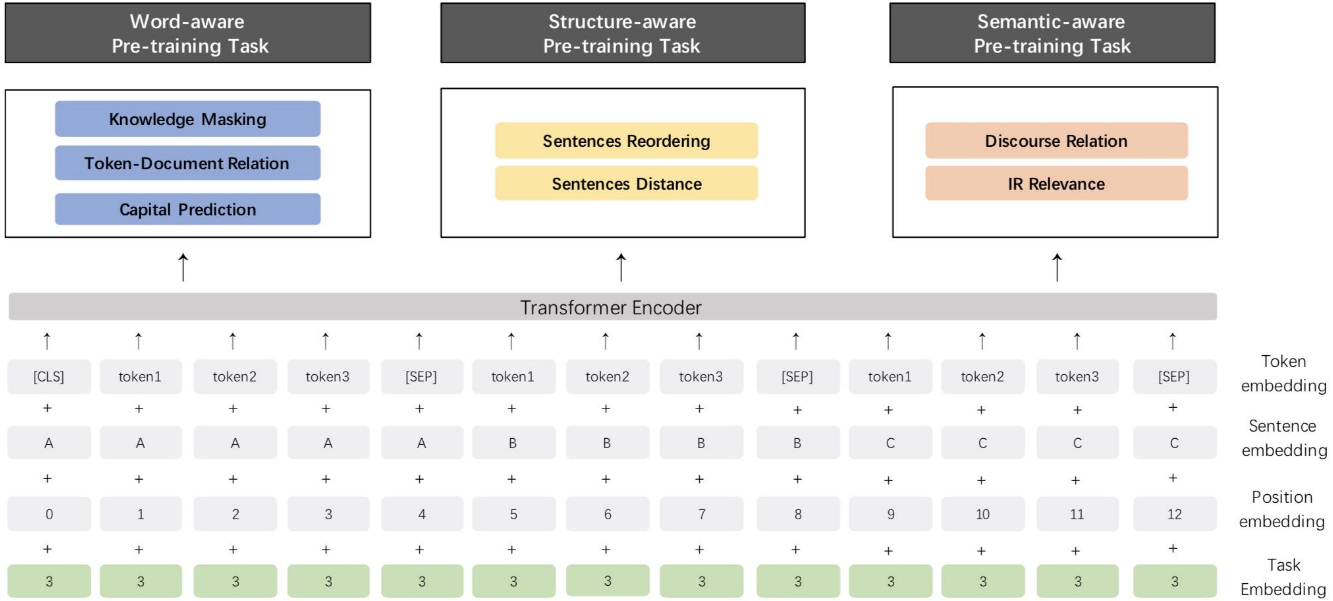
Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian,
Hua Wu, Haifeng Wang

Baidu Inc. (November 2019)

One More Idea



Multi-task Learning



Word-aware Pre-training Task

- Knowledge Masking
- Token-Document Relation
- Capital Prediction

Structure-aware Pre-training Task

- Sentences Reordering
- Sentences Distance

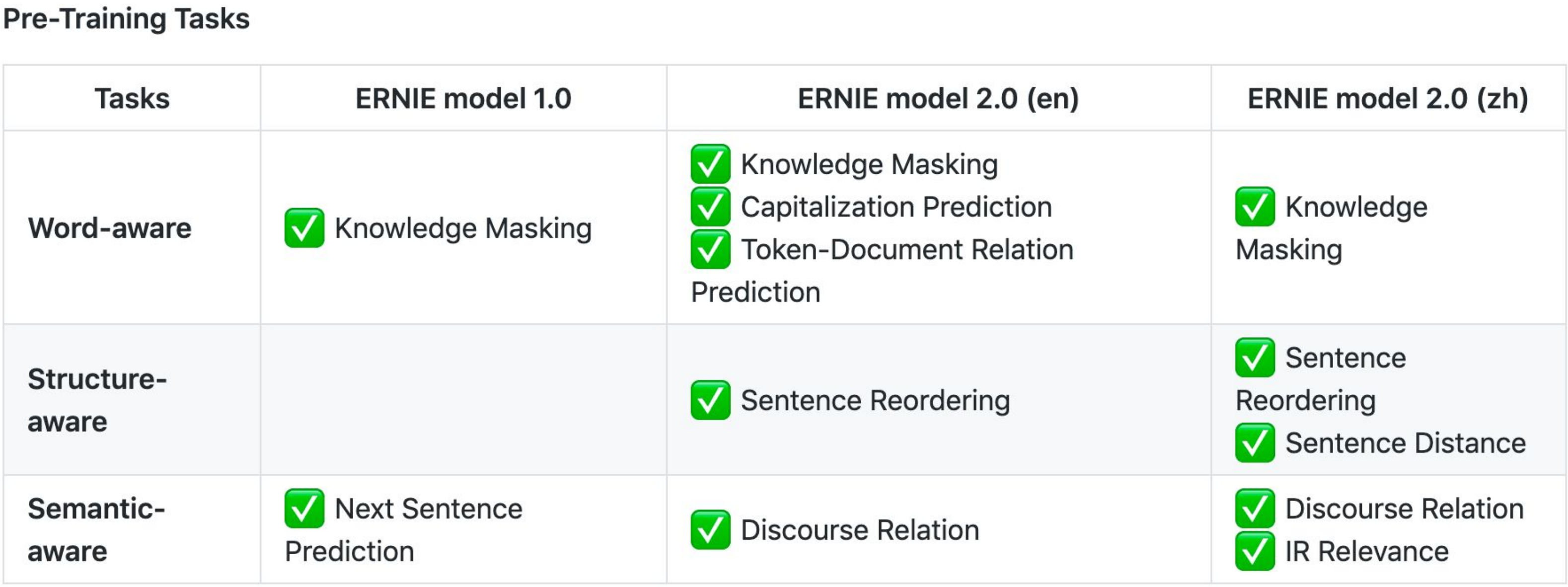
Semantic-aware Pre-training Task

- Discourse Relation
- IR Relevance

Transformer Encoder

[CLS]	token1	token2	token3	[SEP]	token1	token2	token3	[SEP]	token1	token2	token3	[SEP]	Token embedding
+	+	+	+	+	+	+	+	+	+	+	+	+	
A	A	A	A	A	B	B	B	B	C	C	C	C	Sentence embedding
+	+	+	+	+	+	+	+	+	+	+	+	+	
0	1	2	3	4	5	6	7	8	9	10	11	12	Position embedding
+	+	+	+	+	+	+	+	+	+	+	+	+	
3	3	3	3	3	3	3	3	3	3	3	3	3	Task Embedding

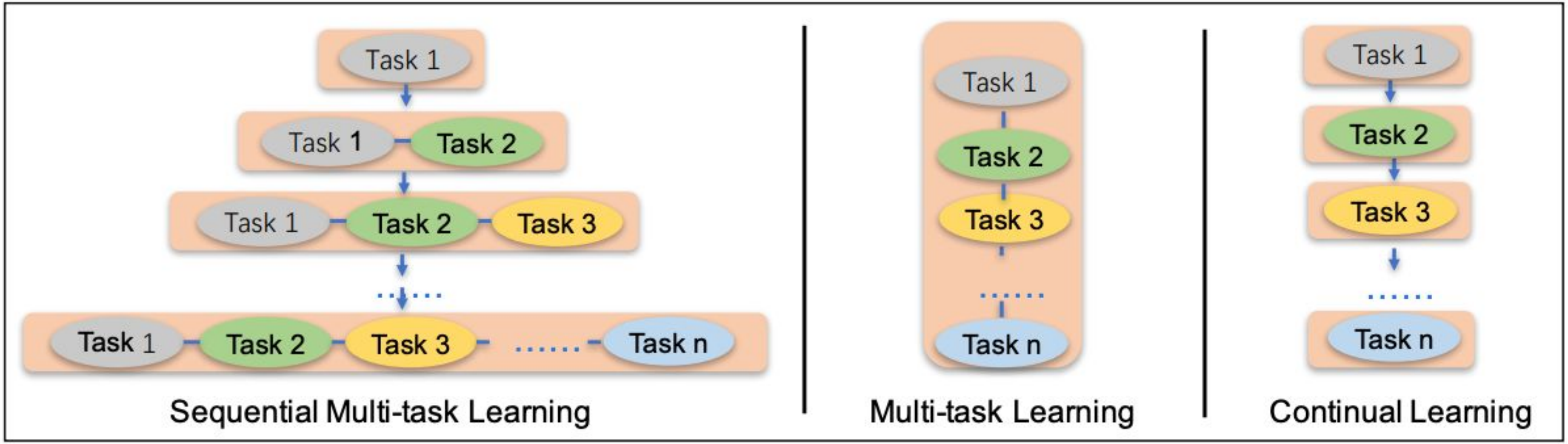
Multi-task Learning



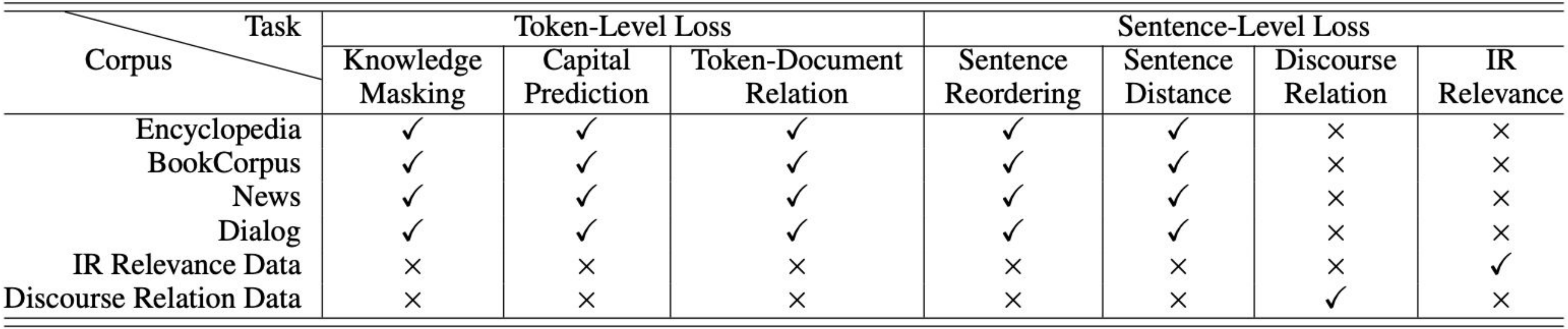
Pre-Training Tasks

Tasks	ERNIE model 1.0	ERNIE model 2.0 (en)	ERNIE model 2.0 (zh)
Word-aware	✅ Knowledge Masking	✅ Knowledge Masking ✅ Capitalization Prediction ✅ Token-Document Relation Prediction	✅ Knowledge Masking
Structure-aware		✅ Sentence Reordering	✅ Sentence Reordering ✅ Sentence Distance
Semantic-aware	✅ Next Sentence Prediction	✅ Discourse Relation	✅ Discourse Relation ✅ IR Relevance

Sequential Multi-task Learning

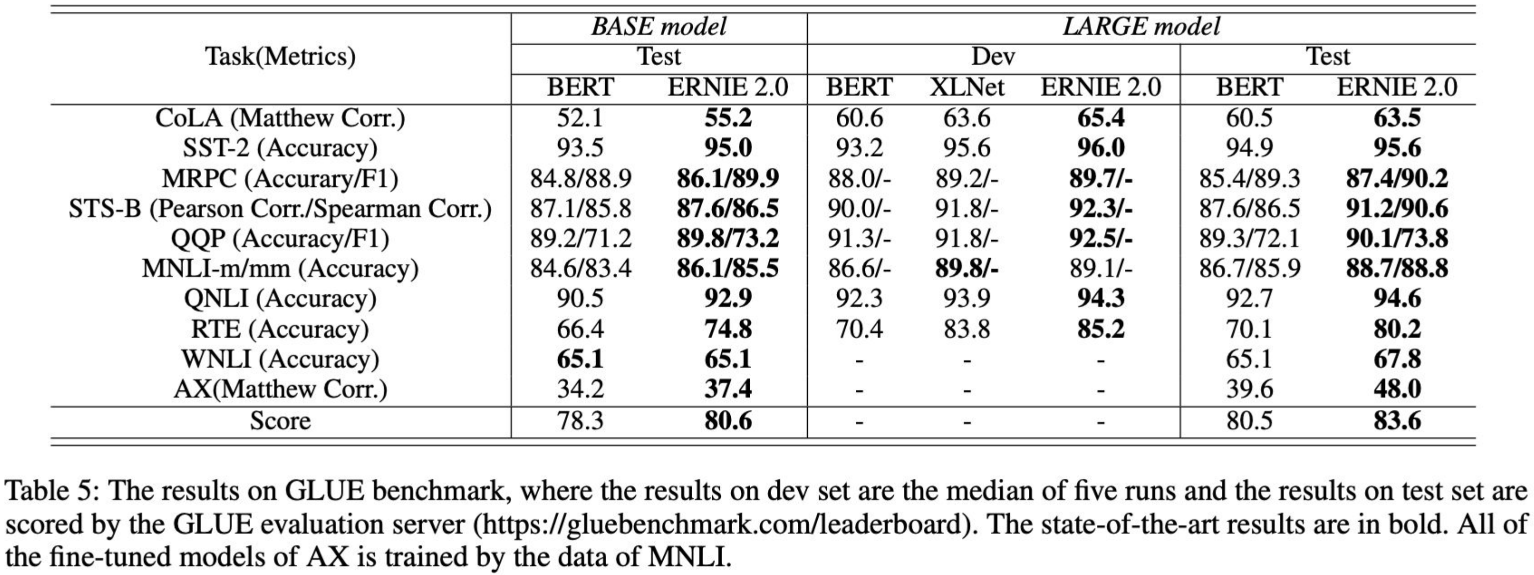


Losses and Data



Task Corpus	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	×	×	×	×	×	✓	×

Results: GLUE



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.

Thank you for your attention