NSP-BERT:

A Prompt-based Few-Shot Learner Through an Original Pre-training Task--Next Sentence Prediction

Advisor : Jia-Ling, Koh

Speaker : Ting-I, Weng

Source : ACL'22

Date : 2023/03/14

Outline

- **Introduction**
- **Definition**
- **Method**
- **Experiment**
- **Conclusion**

Introduction

- Why
- How
- Application
- Improvement

Introduction

- Why?
  - GPT3 has too many parameters, and there is not enough hardware equipment to train the model
  - The length of the prompt is fixed

- How?
  - Use prompts instead of finetune
  - Using the Next Sentence preidiciton, one of the BERT pre-training tasks
  - Bert's NSP can be used to determine whether two sentences are connected

$$\mathbf{x}_{input} = \texttt{[CLS]}\ \mathbf{x}\ \text{It was}\ \texttt{[MASK].[EOS]}$$

$$\mathbf{x}_{input} = \texttt{[CLS]}\ \mathbf{x}\ \text{这是}\ \texttt{[MASK][MASK]}\text{新闻}\texttt{.[EOS]}$$

Introduction

- Application?
  - Suitable for Chinese NLP tasks
    - FIFA = 國際足球總會
    - Sport = 運動
  - Single sentence classification, sentence pair classification...

- Improvement?
  - Significant improvement on single sentence classification tasks compared to GPT-zero and PET-zero
  - Not limited to entity descriptions of different lengths

$$\mathbf{x}_{input} = \texttt{[CLS]}\ \mathbf{x}\ \text{It was}\ \texttt{[MASK].[EOS]}$$

$$\mathbf{x}_{input} = \texttt{[CLS]}\ \mathbf{x}\ \text{这是}\ \texttt{[MASK][MASK]}\text{新闻}\texttt{.[EOS]}$$

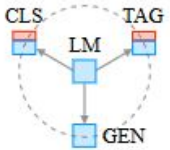
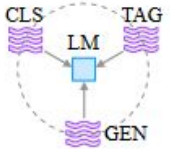
Definition

- Prompt Learning
- Token/Sentence level Prompt Learning

- ref: Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing (arXiv:2107.13586)
- ref: https://www.youtube.com/watch?v=mol6U__9520

Definition:Prompt Learning

Time

Paradigm	Engineering	Task Relation	e.g.
a. Fully Supervised Learning (Non-Neural Network)	Features (e.g. word identity, part-of-speech, sentence length)		SVM、boosting...
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)		CNN、RNN...
c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)		BERT-like
d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)		PET、NSP-BERT...

- ref: Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing (arXiv:2107.13586)

Definition:Prompt Learning

Traditional Formulation V.S Prompt Formulation

Input: x = "I love this movie"

Predicting: y = Positive

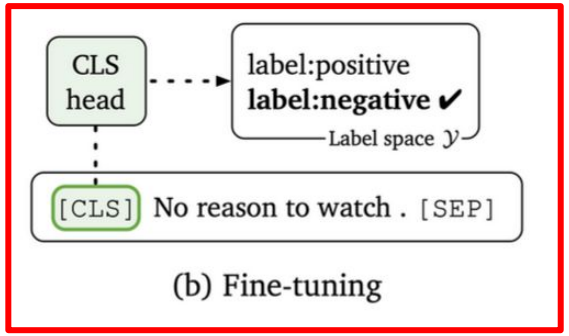
Input: x = "I love this movie"

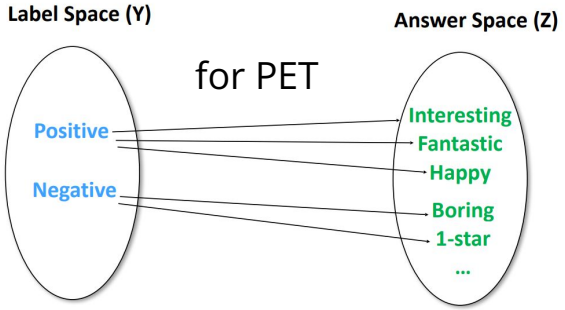
Template: [x] Overall, it was a [z] movie

Prompting: x' = "I love this movie. Overall it was a [z] movie."

Predicting: x' = "I love this movie. Overall it was a fantastic movie."

Mapping (answer -> label):
fantastic => Positive

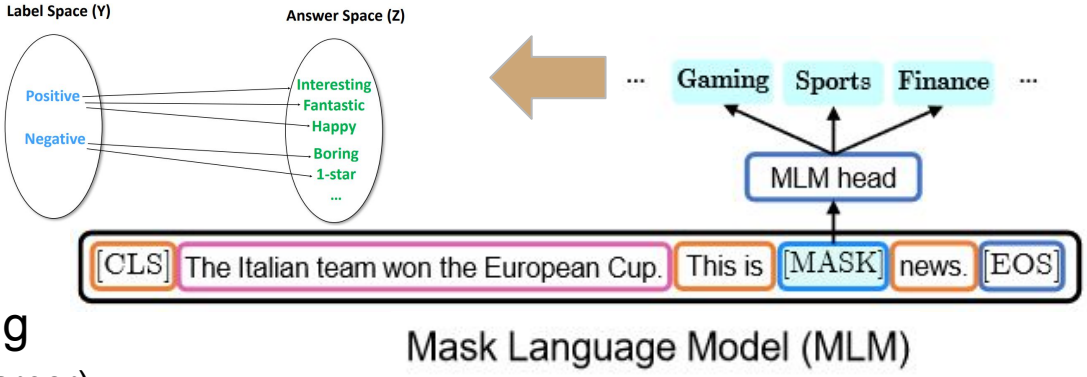


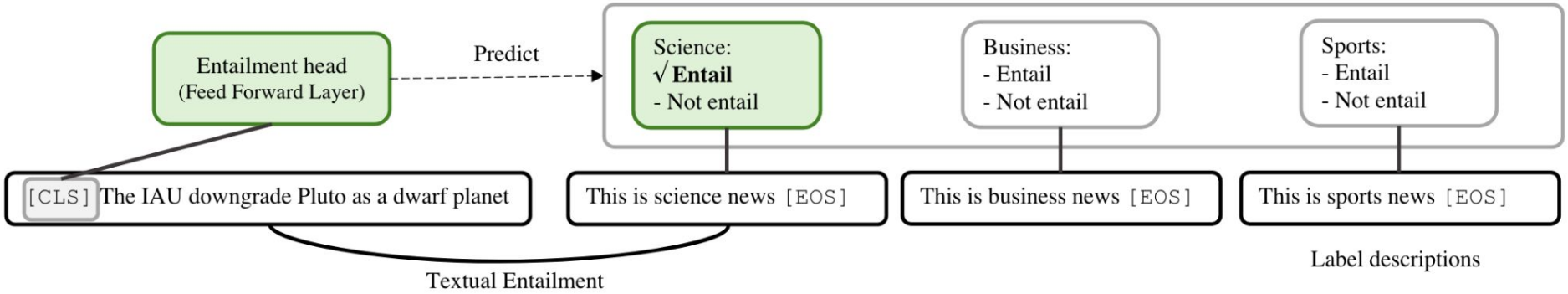


- ref: Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing (arXiv:2107.13586)

Definition:Token/Sentence Level

- Token level prompt-learning
  - MLM
- Sentence level prompt-learning
  - EFL(Entailment as Few-Shot Learner)





Definition:Token/Sentence Level

Difference

	Unit	granularity(粒度)	contextual
token level(MLM)	token	fine-grained	Lack of consideration
Sentence level(NSP)	sentence	Coarse grained	better capture

Method

Prediction
-BERT
g
es-Contrast
Contrast

ref: https://jalammar.github.io/illustrated-bert/

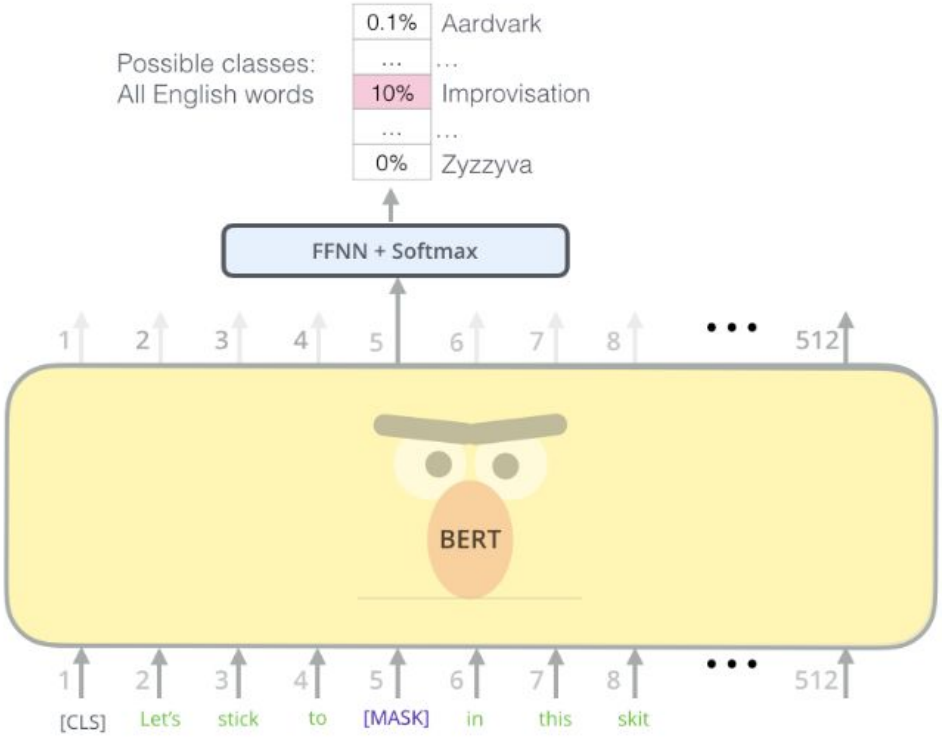
Method: Problem of MLM

$$\mathbf{x}_{input} = \texttt{[CLS]}\ \mathbf{x}\ \texttt{It was [MASK].[EOS]}$$

x = FIFA unveils biennial World Cup plan, UEFA threatens boycott

$$\mathbf{x}_{input} = \texttt{[CLS]}\ \mathbf{x}\ \texttt{这是 [MASK][MASK]新闻.[EOS]}$$

x = 國際足球總會公佈兩年一度的世界杯計劃，歐洲足球總會威脅抵制



ref: https://jalammar.github.io/illustrated-bert/

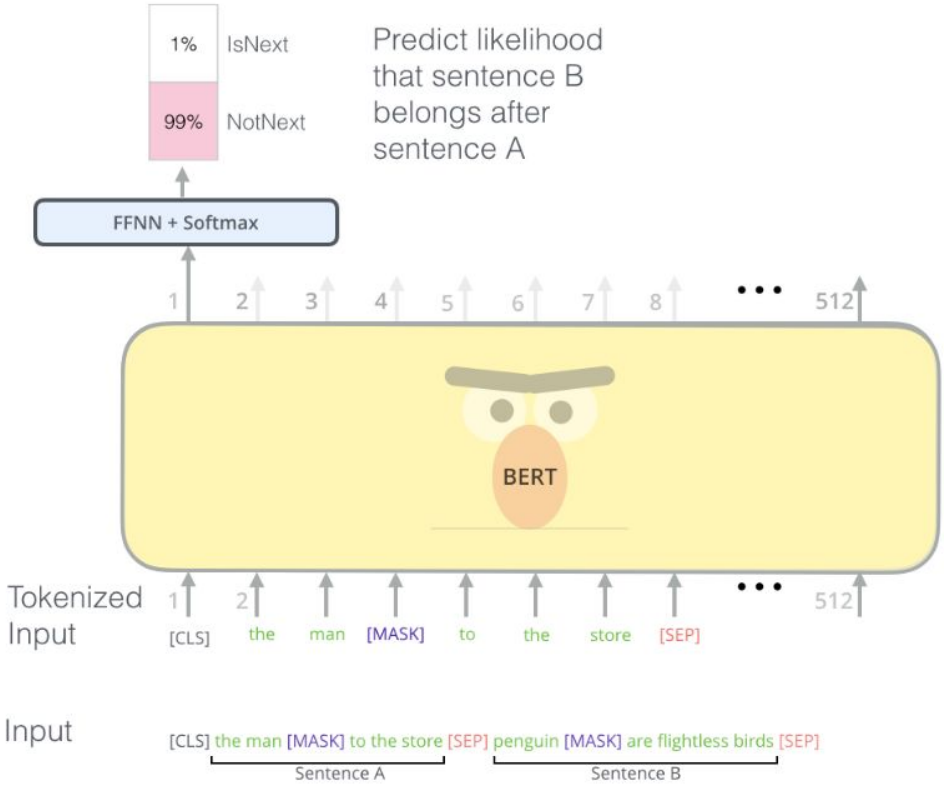
Method: Next Sentence Prediction

Next Sentence Prediction(NSP)

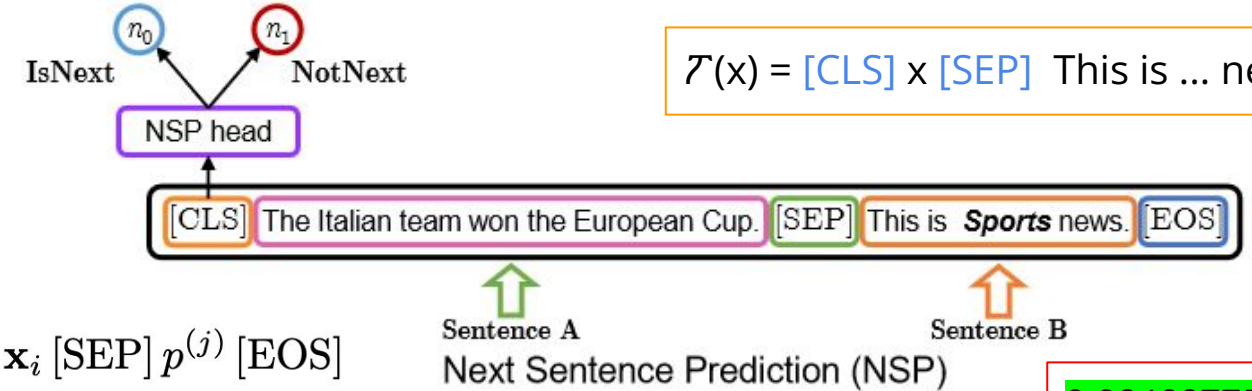
- Determine whether two sentences are connected

$$\mathbf{s} = \mathbf{W}_{\text{nsp}}(\tanh\left(\mathbf{W}\mathbf{h}_{[\text{CLS}]} + \mathbf{b}\right))^2$$

$$q_{\mathcal{M}}(n_k|\mathbf{x}_i) = \frac{\exp s(n_k|\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)})}{\sum_n \exp s(n|\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)})}$$



Method: Prompts in NSP-BERT



$\mathcal{T}$(x) = [CLS] x [SEP] This is … news. [EOS]

$$\mathbf{x}_{input} = [\text{CLS}]\, \mathbf{x}_i\, [\text{SEP}]\, p^{(j)}\, [\text{EOS}]$$

0.98468775 , 0.01531232

$\mathcal{T}$(x) =[CLS] 國際足球總會公佈兩年一度的世界杯計劃，歐洲足球總會威脅抵制 [SEP] 這是一篇 {運動} 新聞 [EOS]

0.67155135 , 0.3284486

$\mathcal{T}$(x) =[CLS] 國際足球總會公佈兩年一度的世界杯計劃，歐洲足球總會威脅抵制 [SEP] 這是一篇 {科技} 新聞 [EOS]

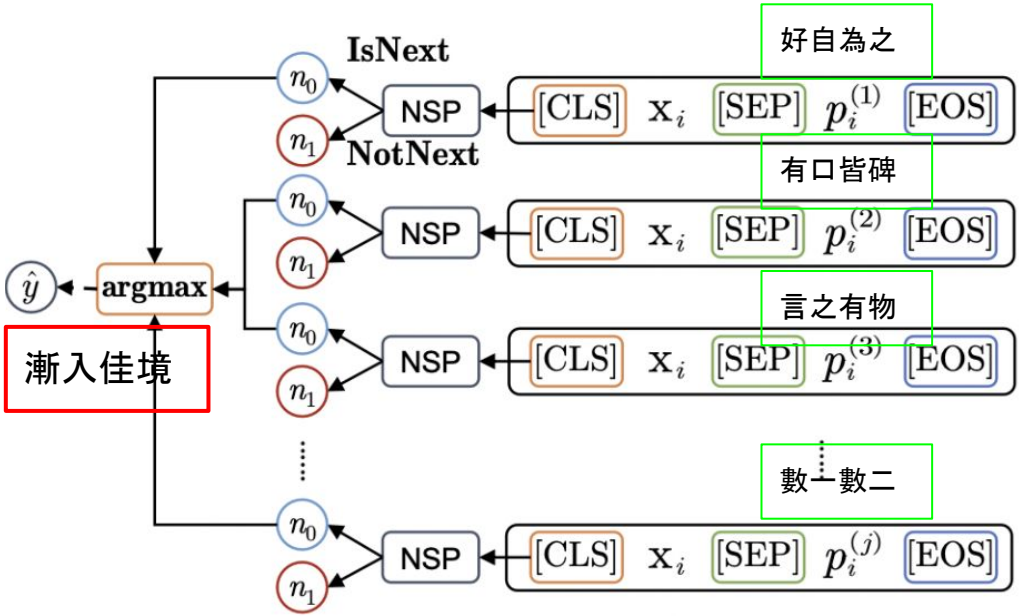
0.4699689 ,0.5300311

$\mathcal{T}$(x) =[CLS] 國際足球總會公佈兩年一度的世界杯計劃，歐洲足球總會威脅抵制 [SEP] 這是一篇 {教育} 新聞 [EOS]

ref: https://blog.51cto.com/u_15919249/5962236

Method: Answer Mapping

Candidates-Contrast



x = 最後廣東新援湯普森在這場也是被安排了首發出場,相比前一場,本場比賽湯普森的表現有所進步,砍下了8分12籃板,可以說湯普森的#idiom#是廣東所期待的,畢竟現在沒了馬尚,季后賽單靠威姆斯一人難免會讓他體力不支,而如果湯普森能在威姆斯休...

candidates: ["好自為之", "有口皆碑", "言之有物", "數一數二", "史無前例", "漸入佳境", "半懂不懂"]

answer": 5

$$q(y_i^{(j)}|\mathbf{x}_i) \propto q_{\mathcal{M}}(n = \texttt{IsNext}|\mathbf{x}_i, p_i^{(j)})$$

正比於

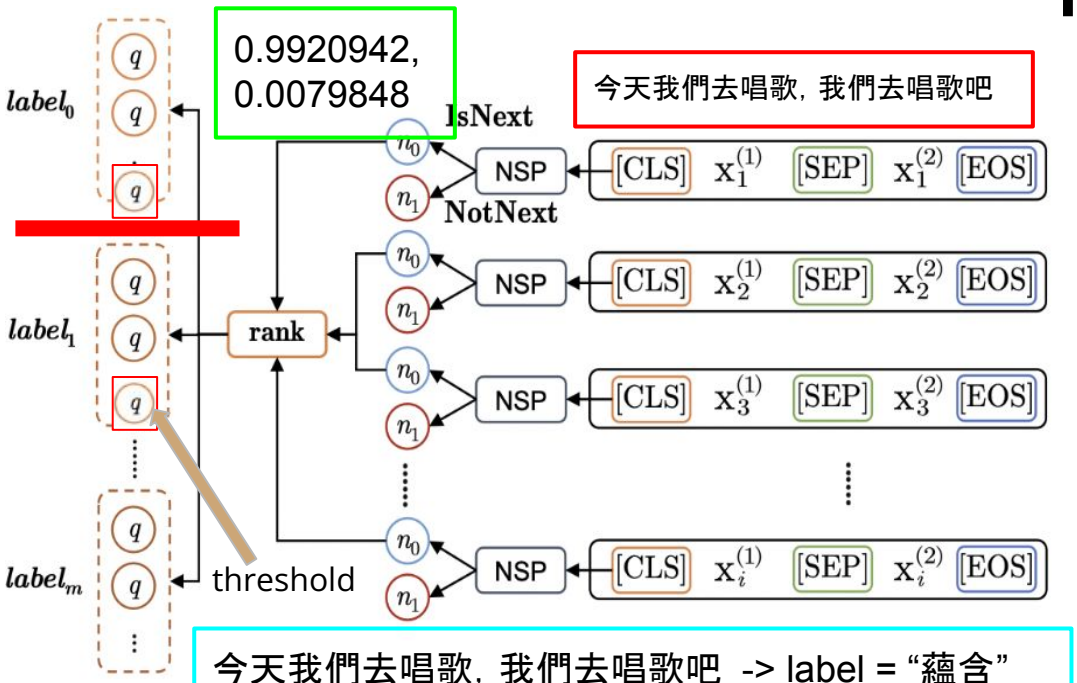
candidates

$$\hat{y}_i = \arg\max_j q(y_i^{(j)}|\mathbf{x}_i)$$
$$= \arg\max_j q_{\mathcal{M}}(n = \texttt{IsNext}|\mathbf{x}_i, p_i^{(j)})$$

ref: https://blog.51cto.com/u_15919249/5962236

Method: Answer Mapping

sentence pair
Samples-Contrast



今天我們去唱歌，我們去唱歌吧 -> label = “蘊含”
面積公式是什麼，Be動詞有哪些 -> label = “無蘊含"

(2, 0.9920942), (9, 0.9972888), (8, 0.9990614),....

Algorithm 1 Samples-Contrast Answer Mapping

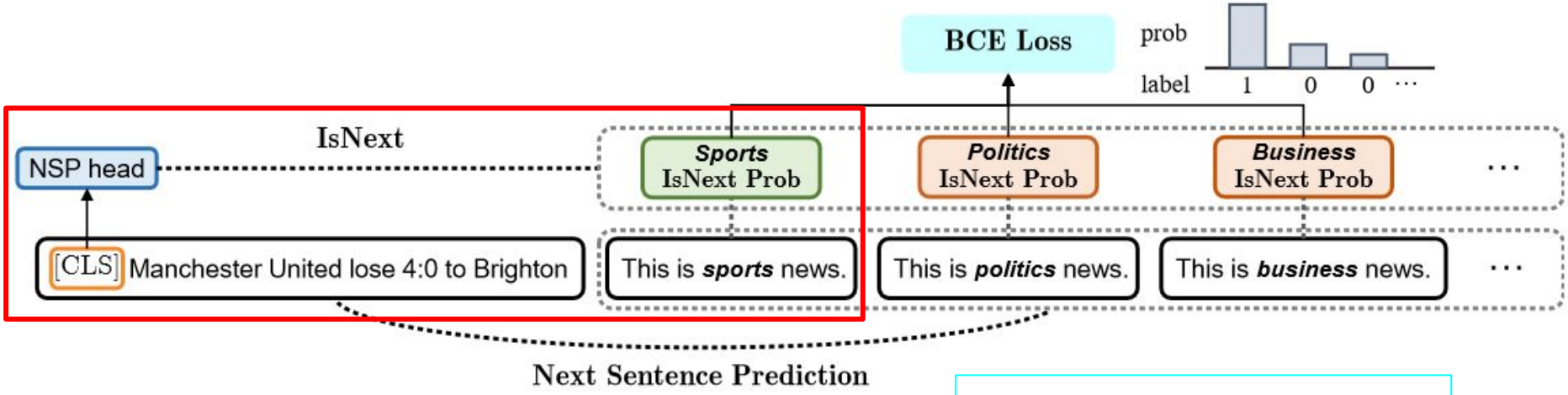
Input: Test set $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}$, where $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)})$, Oder $o \in$ {“ascending”, “descending”}, distribution of labels d, batch size bs.

Output: $\{\mathbf{x}_i, \hat{y}_i\}_{i=1}^{N}$

for $i = 1, ..., N$ **do**
  $q_i \leftarrow q_{\mathcal{M}}(n = \text{IsNext}|\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)})$
end for
$\{\mathcal{B}_j\}_{j=1}^{\lceil \frac{N}{bs} \rceil} \leftarrow$ **divide** $(\mathcal{D}, bs)$
for $j = 1, ..., \lceil \frac{N}{bs} \rceil$ **do**
  $\mathcal{B}'_j = \{\mathbf{x}_{r(1)}, ..., \mathbf{x}_{r(bs)}\} \leftarrow$ **rank**$(\mathcal{B}_j, q_i, o)$
  $\{B_m\}_{m=1}^{M} \leftarrow$ **divide** $(\mathcal{B}'_j, d)$
  for $i = 1, ..., bs$ **do**
    $\hat{y}_i \leftarrow m$ **where** $\mathbf{x}_i \in B_m$
  end for
end for

Method: NSP-tuning

1. Building Instances
2. Loss function



single sentence prediction

Experiment

- Datasets
- Few/zero shot learning results
- Impact of pre-training corpus
- Different Batch Size
- Ablation Studies

Experiment

- Compare MLM's PET 、sentence level EFL(Entailment as Few shot learner) and NSP-BERT
- Based on two models of roberta and bert, used in different corpora, compared with PET
- Observe different batch sizes
- Ablation experiment

Datasets

corpus	SST	MR	CR	MPQA	Subj	Yahoo	AGNews
Source	movie Reviews	movie Reviews	E-commerce Reviews	World Press	Movie Reviews	Yahoo	Web
Task Type	sentiment analysis	sentiment analysis	sentiment analysis	Opinion Polarity	Subjectivity(主觀)	Question Classification	News Topic Classification

corpus	EPRSTMT	TNEWS	CSLDCP	IFLYTEK
Source	E-commerce Reviews	News Title	Academic CNKI	App Description
Task Type	sentiment analysis	Short Text Classification	Long Text Classification	Long Text Classification

TNEWS：股票全天波动-1%左右，尾盘三分钟拉回，第二天低开，主力意图是吸筹还是洗盘？news_stock

IFLYTEK：《香港远足路线》是一个香港行山手机程。旅游资讯

Experiment: Zero/few-shot learning

- task : single sentence classification
- metric : accuracy
- model
  - English : BERT- large
  - Chinese : UER's Chinese BERT-base
- baseline : PET、EFL

TNEWS(K
"sentence"
"keywords'

Corpus	#Train	#Test	$\|\mathcal{Y}\|$
SST-2	6,920	872	2
MR	8,662	2,000	2
CR	1,775	2,000	2

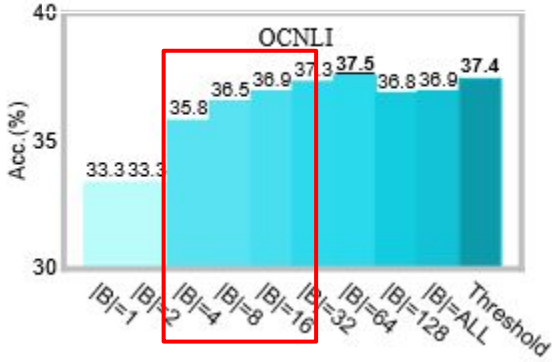
		English Tasks							Chinese Tasks			
		SST-2	MR	CR	MPQA	Subj	Yahoo!	AGNews	EPR.	TNEWS(K)	CSLDCP	IFLY.
Full	Majority	*50.9*	*50.0*	*50.0*	*50.0*	*50.0*	*10.0*	*25.0*	*50.0*	*6.7*	*1.5*	*0.8*
	Fine-Tuning	*93.6*	*89.0*	*89.3*	*89.3*	*97.0*	*76.5*	*94.7*	*90.0†*	*71.0†*	*68.0†*	*66.0†*
Zero	PET	67.6	65.3	**61.2**	**63.9**	**61.0**	25.6	54.5	60.7	28.0 / 35.6	22.4	34.8
	NSP-BERT	**75.6**	**74.4**	59.4	59.9	53.9	**47.0**	**77.5**	**86.9**	**51.9 / 57.0**	**47.6**	**41.6**
Few	Fine-tuning	77.9±5.9	68.0±9.4	79.1±8.9	65.2±6.3	**89.7±1.1**	61.8±1.5	82.4±1.2	78.7±5.8	51.1±1.1 / 58.0±1.4	51.7±2.1	45.1±2.2
	PET	86.0±1.6	80.0±1.6	**88.9±0.6**	83.3±2.4	86.2±1.5	64.3±1.3	84.2±0.8	82.5±2.0	54.7±1.1 / 61.2±0.9	52.6±1.2	45.9±2.1
	EFL w/ PT	**86.9±1.8**	**80.6±1.2**	88.1±0.9	**86.1±0.7**	86.0±3.3	63.0±1.2	83.8±1.3	84.8±1.6	53.2±1.5 / 59.2±1.6	52.0±1.6	47.9±1.5
	EFL w/o PT	81.2±5.1	76.1±9.1	79.2±4.0	79.1±1.6	75.1±9.4	60.8±4.2	84.6±0.7	84.6±2.1	54.7±1.3 / 60.3±1.7	53.8±0.9	49.5±1.2
	NSP-BERT	86.8±1.3	80.5±1.5	86.0±2.2	83.9±1.1	86.4±1.8	**64.5±0.5**	**85.9±0.8**	**87.7±0.7**	**55.7±1.0 / 61.6±0.9**	**55.0±1.5**	**49.5±1.1**

Experiment: Pre-training corpus

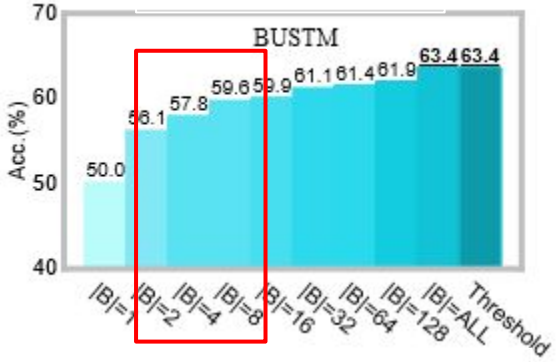
		Model	Corpus	English Tasks						
				SST-2	MR	CR	MPQA	Subj	Yahoo!	AGNews
Zero	PET	RoBERTa	$\mathcal{C}_{\text{B}}$	81.2	75.6	76.6	63.3	**63.6**	18.7	47.8
			$\mathcal{C}_{\text{R}}$	**83.6**	**80.8**	**79.5**	**67.6**	53.6	**25.6**	**54.5**
		BERT	$\mathcal{C}_{\text{B}}$	67.6	65.3	61.2	63.9	**61.0**	25.6	**54.5**
			$\mathcal{C}_{\text{B+Mix5}}$	**75.0**	**70.1**	**67.4**	**64.2**	55.3	**28.5**	38.4
	NSP-BERT	BERT	$\mathcal{C}_{\text{B}}$	75.6	74.4	59.4	59.9	**53.9**	47.0	**77.5**
			$\mathcal{C}_{\text{B+Mix5}}$	**81.2**	**78.3**	**76.9**	**72.4**	53.0	**56.8**	75.8
Few	PET	RoBERTa	$\mathcal{C}_{\text{B}}$	88.6±1.5	83.9±0.8	87.8±0.7	82.0±1.1	82.8±5.6	65.2±1.3	86.0±0.4
			$\mathcal{C}_{\text{R}}$	**91.7±0.6**	**88.0±0.5**	**91.5±0.9**	**85.6±2.1**	**87.8±2.2**	**68.9±1.0**	**87.8±0.9**
		BERT	$\mathcal{C}_{\text{B}}$	85.3±1.7	80.3±2.1	89.2±0.3	83.3±2.4	85.4±1.9	64.3±1.3	84.0±1.0
			$\mathcal{C}_{\text{B+Mix5}}$	**87.6±0.9**	**85.0±0.8**	**89.6±0.8**	**85.0±1.7**	**90.5±1.2**	**68.4±0.7**	**87.8±0.6**
	NSP-BERT	BERT	$\mathcal{C}_{\text{B}}$	86.7±2.1	80.3±1.8	86.7±1.7	83.9±1.1	86.6±0.9	64.5±0.5	85.9±0.8
			$\mathcal{C}_{\text{B+Mix5}}$	**89.4±0.7**	**83.3±1.1**	**88.7±1.0**	**85.3±1.0**	**92.1±1.1**	**68.3±1.3**	**87.6±0.5**

Experiment: Different Batch Size

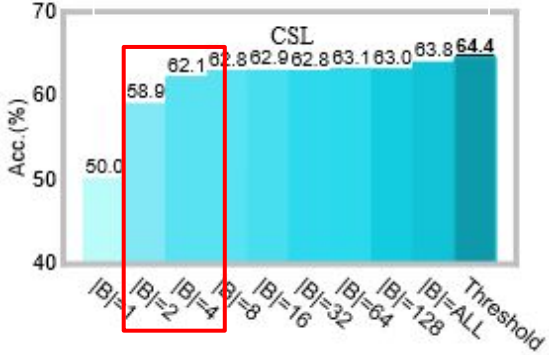
natural language inference



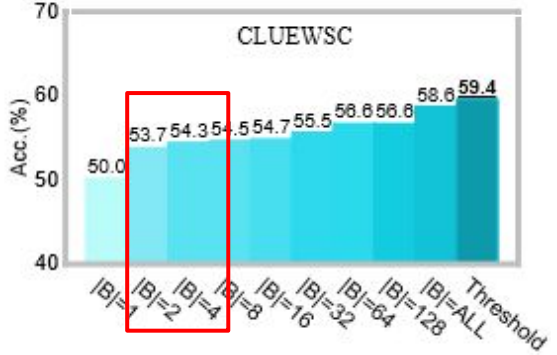
text matching



keyword recognition



Chinese idiom cloze test



Experiment: Ablation Studies

- Model : BERT-Large
- Tasks : English single sentence classification

	SST-2	MR	CR	MPQA
NSP-BERT	86.8±1.3	**80.5±1.5**	**86.0±2.2**	**83.9±1.1**
coupled→decouple	**86.8±1.2**	78.9±2.3	85.8±9.5	81.5±5.8
BCE→softmax	83.8±5.0	76.4±6.4	80.5±10.0	73.3±9.5
w/o NSP head	83.8±6.5	74.3±9.2	79.0±8.1	73.2±10.1
linear head+softmax	80.2±7.6	71.9±12.3	82.6±6.7	73.8±11.1

Conclusion

Conclusion

- NSP can also be the same as MLM, suitable for few shot, zero shot learning
- Rethink the role of sentence level in NLP tasks
- The size of the model or the pre-trained corpus is the upper limit of the ability of few shot learning to determine the model

Thanks for your listening