Multi-Task Deep Neural Networks for Natural Language Understanding

QRAFT | AXE

GUIJIN SON

YOnsei natural language Generation and Understanding paper Reading sTudy

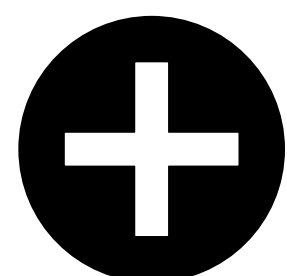# Natural Language Understanding

Multi-Task Deep Neural Networks for Natural Language Understanding
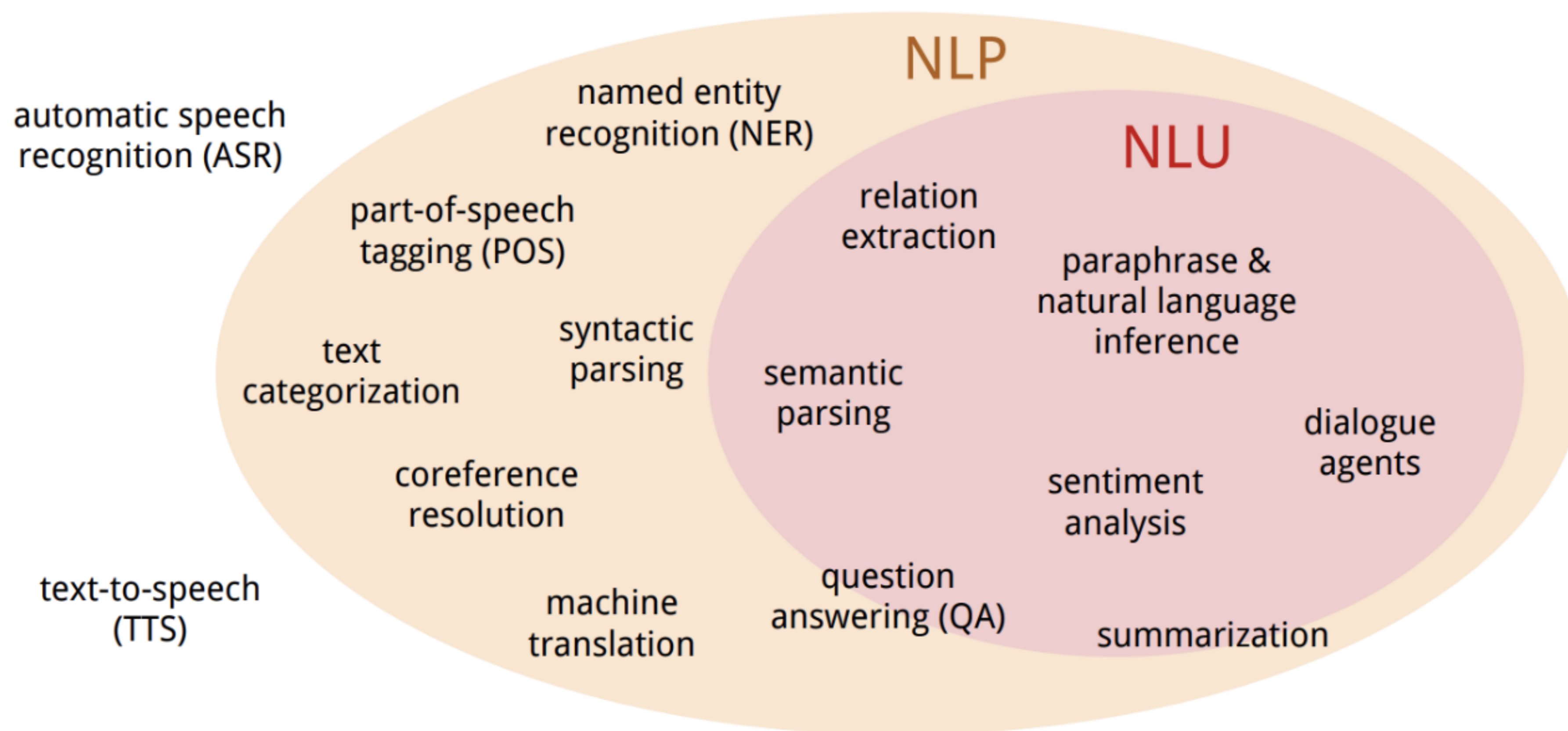
Excelling in NLU Tasks
Multi-Task Deep Neural Networks for Natural Language Understanding
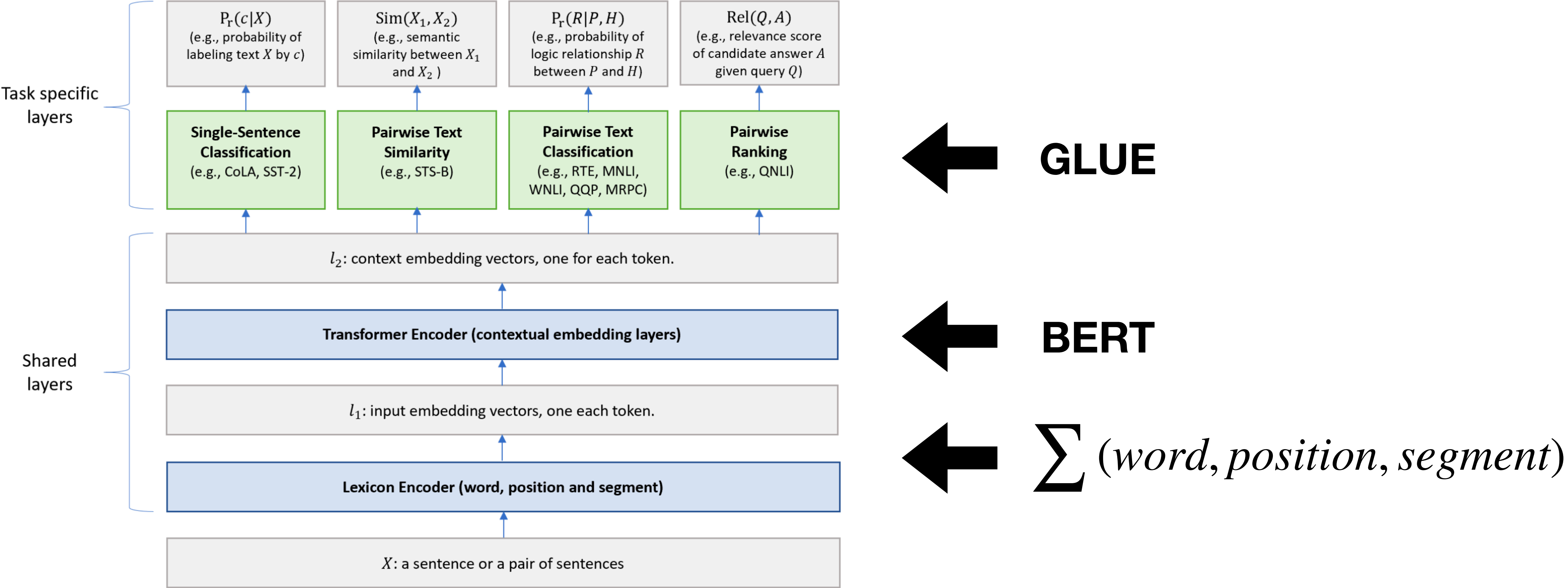
**Language Model Pretraining**

➕

Multi-Task Learning

=   MT-DNN

Model Architecture
Multi-Task Deep Neural Networks for Natural Language Understanding

Task specific layers

$\mathrm{P_r}(c|X)$
(e.g., probability of labeling text X by c)

$\mathrm{Sim}(X_1, X_2)$
(e.g., semantic similarity between X_1 and X_2)

$\mathrm{P_r}(R|P, H)$
(e.g., probability of logic relationship R between P and H)

$\mathrm{Rel}(Q, A)$
(e.g., relevance score of candidate answer A given query Q)

Single-Sentence Classification
(e.g., CoLA, SST-2)

Pairwise Text Similarity
(e.g., STS-B)

Pairwise Text Classification
(e.g., RTE, MNLI, WNLI, QQP, MRPC)

Pairwise Ranking
(e.g., QNLI)

← **GLUE**

Shared layers

l_2: context embedding vectors, one for each token.

Transformer Encoder (contextual embedding layers)

← **BERT**

l_1: input embedding vectors, one each token.

Lexicon Encoder (word, position and segment)

← $\sum (word, position, segment)$

X: a sentence or a pair of sentences

Training Stages

Multi-Task Deep Neural Networks for Natural Language Understanding

Pretraining

Masked Language Modeling

Next Sentence Prediction

Multi-Task Learning

GLUE Dataset

➡ **Fine-Tuning**
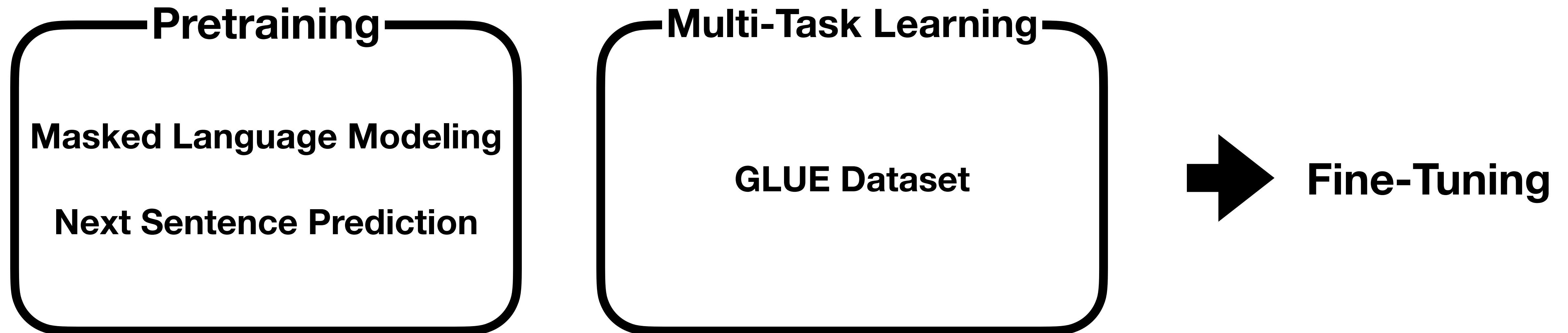
Performance
Multi-Task Deep Neural Networks for Natural Language Understanding

Model	CoLA 8.5k	SST-2 67k	MRPC 3.7k	STS-B 7k	QQP 364k	MNLI-m/mm 393k	QNLI 108k	RTE 2.5k	WNLI 634	AX	Score
BiLSTM+ELMo+Attn [1]	36.0	90.4	84.9/77.9	75.1/73.3	64.8/84.7	76.4/76.1	-	56.8	65.1	26.5	70.5
Singletask Pretrain Transformer [2]	45.4	91.3	82.3/75.7	82.0/80.0	70.3/88.5	82.1/81.4	-	56.0	53.4	29.8	72.8
GPT on STILTs [3]	47.2	93.1	87.7/83.7	85.3/84.8	70.1/88.1	80.8/80.6	-	69.1	65.1	29.4	76.9
BERT$_{LARGE}$[4]	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7/85.9	92.7	70.1	65.1	39.6	80.5
MT-DNN$_{no-fine-tune}$	58.9	94.6	**90.1/86.4**	89.5/88.8	**72.7/89.6**	86.5/85.8	**93.1**	79.1	65.1	39.4	81.7
MT-DNN	**62.5**	**95.6**	**91.1/88.2**	**89.5/88.8**	**72.7/89.6**	**86.7/86.0**	**93.1**	**81.4**	65.1	**40.3**	**82.7**
Human Performance	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0/92.8	91.2	93.6	95.9	-	87.1

Model	MNLI-m/mm	QQP	RTE	QNLI (v1/v2)	MRPC	CoLa	SST-2	STS-B
BERT$_{LARGE}$	86.3/86.2	91.1/88.0	71.1	90.5/92.4	89.5/85.8	61.8	93.5	89.6/89.3
ST-DNN	86.6/86.3	91.3/88.4	72.0	96.1/-	89.7/86.4	-	-	-
MT-DNN	**87.1/86.7**	**91.9/89.2**	**83.4**	**97.4/92.9**	**91.0/87.5**	**63.5**	**94.3**	**90.7/90.6**

Performance
Multi-Task Deep Neural Networks for Natural Language Understanding

(a) SNLI
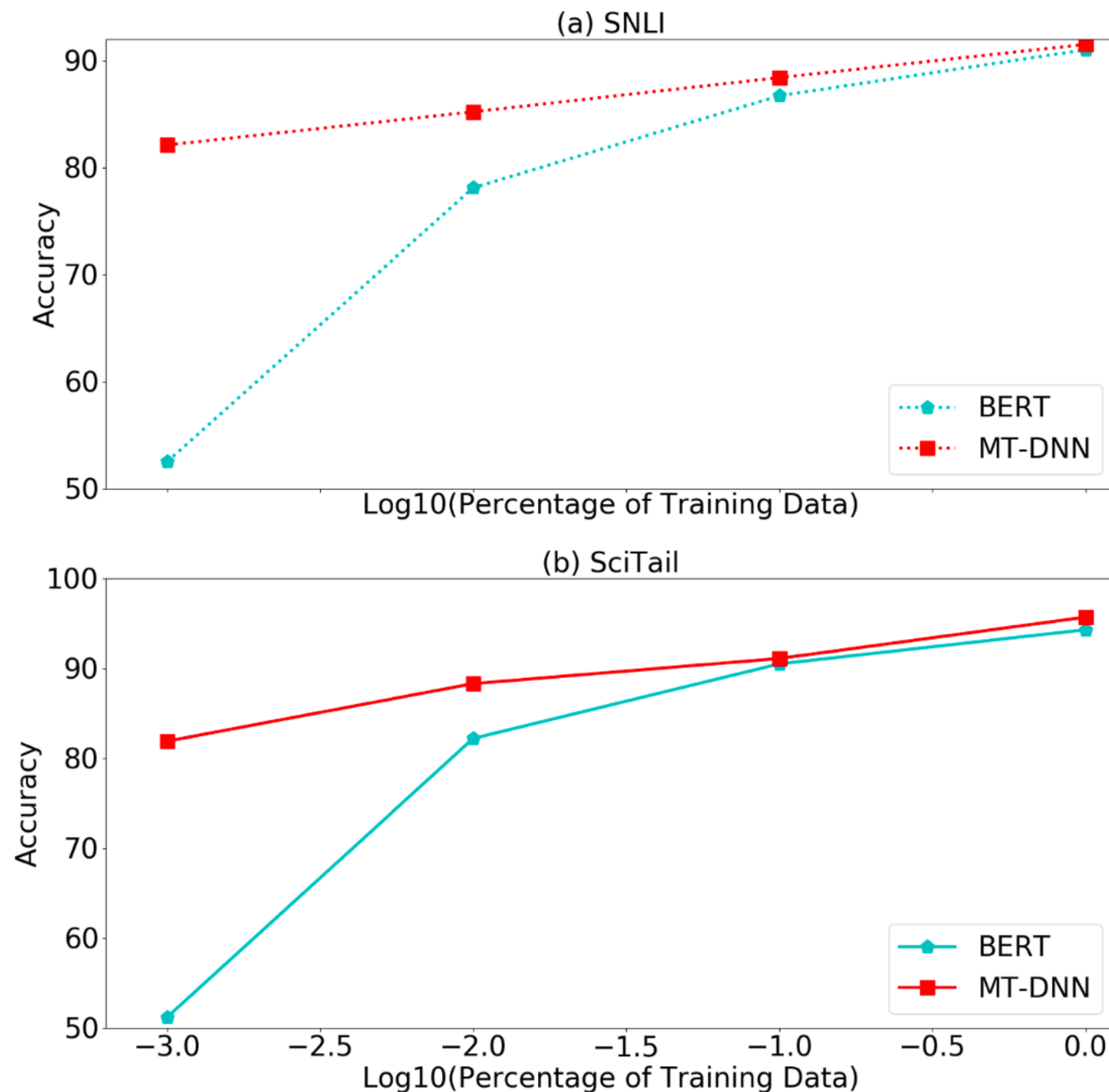
(b) SciTail

Model	0.1%	1%	10%	100%
SNLI Dataset (Dev Accuracy%)				
#Training Data	549	5,493	54,936	549,367
BERT	52.5	78.1	86.7	91.0
MT-DNN	82.1	85.2	88.4	91.5
SciTail Dataset (Dev Accuracy%)				
#Training Data	23	235	2,359	23,596
BERT	51.2	82.2	90.5	94.3
MT-DNN	81.9	88.3	91.1	95.7

Implications

Multi-Task Deep Neural Networks for Natural Language Understanding

- Multi-Task Learning provides an effective way of leveraging supervised data from many related tasks

- The use of multi-task learning profits from a regularization effect via al- leviating overfitting to a specific task, thus making the learned representations universal across tasks.

.