

Dialogue Quality and Nugget Detection for Short Text Conversation (STC-3) based on Hierarchical Multi-Stack Model with Memory Enhance Structure

Team: WIDM

Hsiang-En Cherng and Chia-Hui Chang

National Central University, Taoyuan, Taiwan

seancherng.tw@gmail.com, chiahui@g.ncu.edu.tw

Task Definition

We consider the DQ and ND subtasks for STC-3 using deep learning method. The goal of NQ and DQ subtasks is to extend the one-round STC to multi-round conversation such as customer-helpdesk dialogues. The DQ subtask aims to judge the quality of the whole dialogue using three measures: Task Accomplishment (A-score), Dialogue Effectiveness (E-score) and Customer Satisfaction of the dialogue (S-score). The ND subtask, on the other hand, is to classify if an utterance in a dialogue contains a nugget, which is similar to dialogue act (DA) labeling problem.

Introduction

We compared several DNN models based on a general model with utterance layer, context layer, memory layer and output layer to learn dialogue representation and use gating and attention mechanism at utterance layer and context layer. We report the performance of not just the uploaded model during STC-3 but also the better model with pre-trained BERT sentence embedding. The former used multi-stack CNN with word2vec input for utterance representation, while the latter use pure BERT for utterance representation. Overall, in both DQ and ND subtasks, the new model results in the best performance than NTCIR baseline models.

Dialogue Quality (DQ)

Model

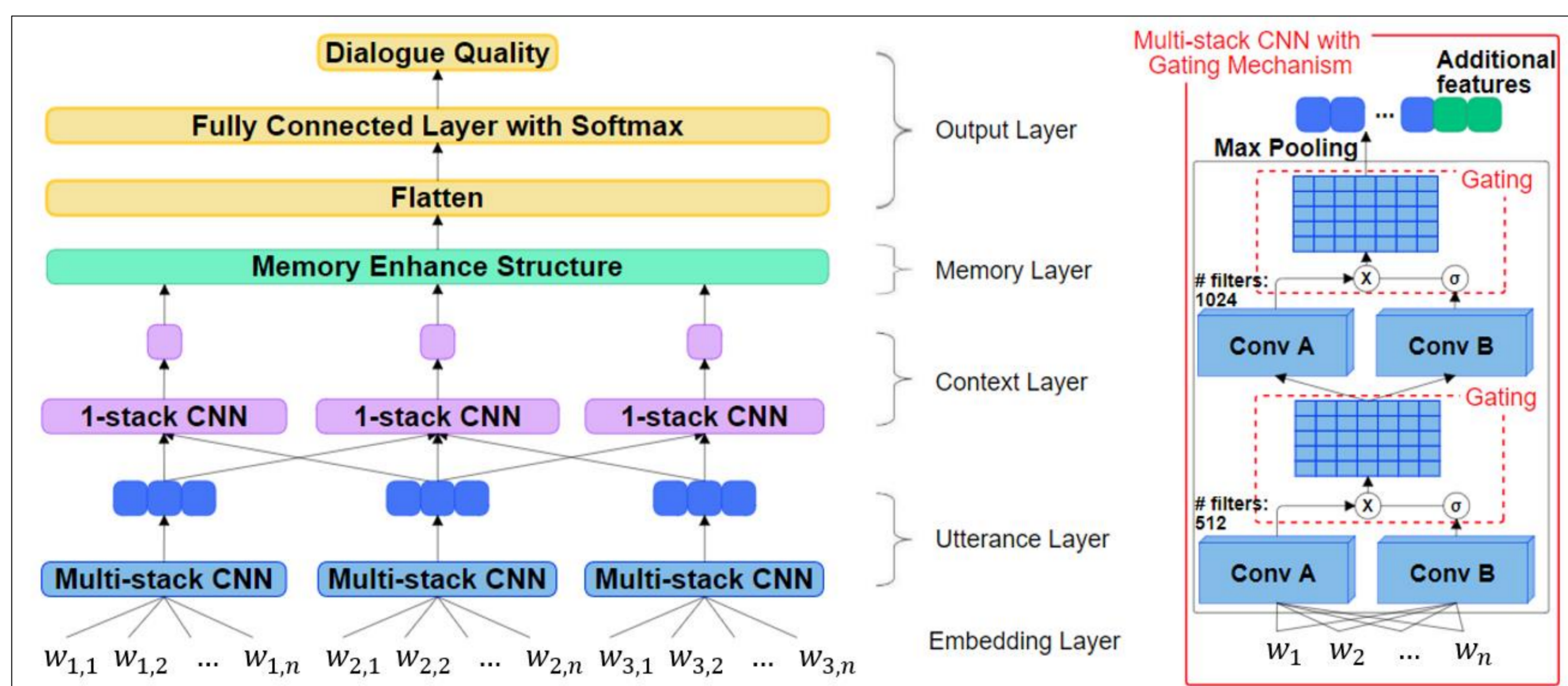


Fig 1. Memory enhance hierarchical gated CNN (MeHGCNN)

- **Embedding Layer:** Word2Vec with 100 dimensions
- **Utterance Layer:** Apply 2-stack gated CNN to learn context information in filter size 2. Additional features are nuggets & speaker features.
- **Context Layer:** 1-stack gated CNN is applied to learn context representation of adjacent utterances
- **Memory Layer:** Memory network structure is applied to capture long-range context features between utterances by self-attention mechanism.
- **Output Layer:** Output the dialogue quality by a simple fully-connected layer with softmax activation function

Experiments

Table 1 shows the performance of DQ subtask in NMD and RSNOD

- **BL-BERT:** Simple BERT without any context or memory layer
- **MeGCBERT:** Replace the embedding layer and utterance layer of MeHGCNN with BERT

Table 1. Performance of DQ subtask

Model	(A-score)		(E-score)		(S-score)	
	NMD	RSNOD	NMD	RSNOD	NMD	RSNOD
BL-uniform	0.1677	0.2478	0.1580	0.2162	0.1987	0.2681
BL-popularity	0.1855	0.2532	0.1950	0.2774	0.1499	0.2326
BL-lstm	0.0896	0.1320	0.0824	0.1220	0.0838	0.1310
BL-BERT	0.0934	0.1379	0.0881	0.1344	0.0842	0.1337
MeHGCNN	0.0862	0.1307	0.0814	0.1225	0.0787	0.1241
MeGCBERT	0.0823	0.1255	0.0791	0.1202	0.0758	0.1245

Table 2 shows the ablation of MeGCBERT. In summary, gating mechanism, memory layer and nugget features all improve A, E and S score

Table 2. Ablation of MEGCBERT

Model	(A-score)		(E-score)		(S-score)	
	NMD	RSNOD	NMD	RSNOD	NMD	RSNOD
MeGCBERT	0.0823	0.1255	0.0791	0.1202	0.0758	0.1245
W/o gating	0.0885	0.1322	0.0813	0.1214	0.0815	0.1289
W/o memory	0.0913	0.1364	0.0808	0.1235	0.0799	0.1273
W/o nuggets	0.0963	0.1388	0.0802	0.1204	0.0774	0.1247

Nugget Detection (ND)

Model

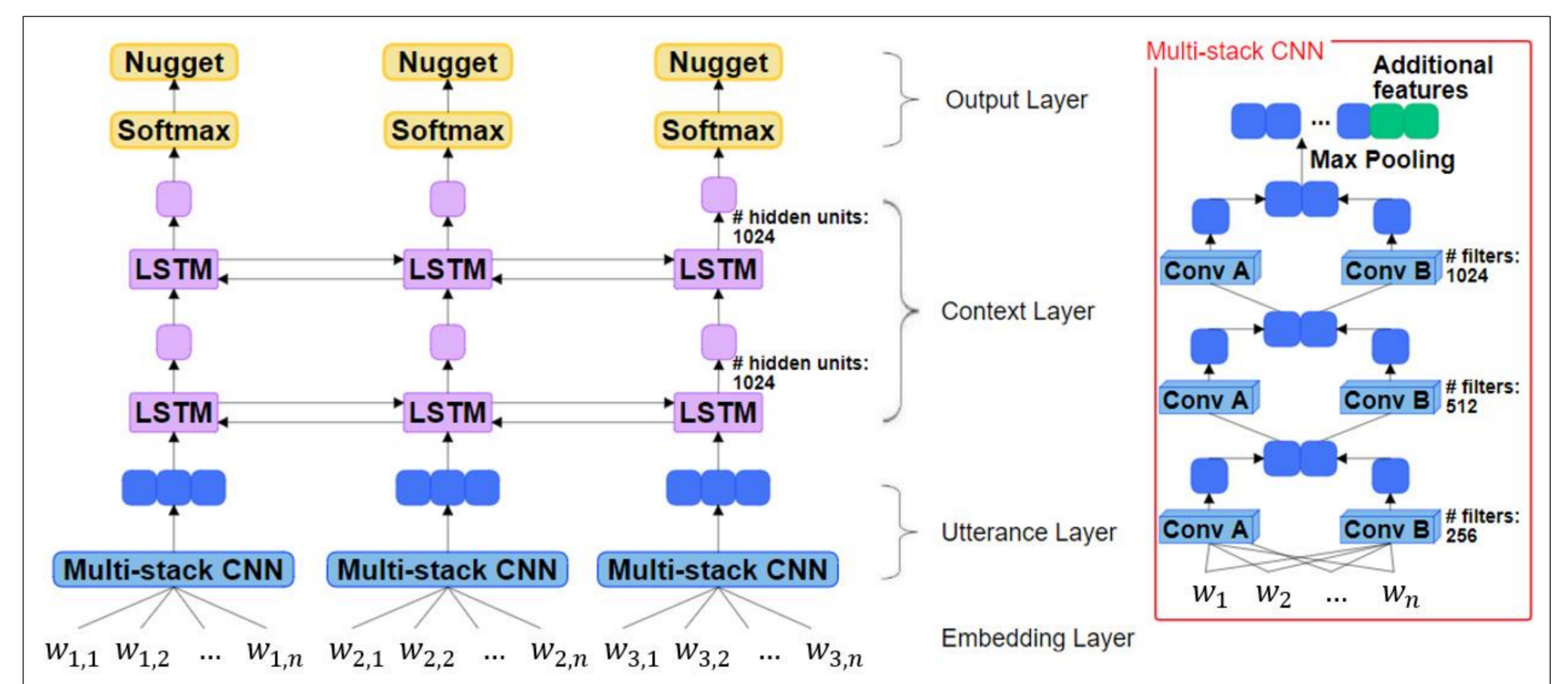


Fig 2. Hierarchical CNN + BI-LSTM (HCNN-LSTM)

- **Embedding Layer:** Word2Vec with 100 dimensions
- **Utterance Layer:** Apply 3-stack CNN by concatenation of two convolutions with filter size 2 and 3 instead of gated CNN to learn utterance representation. Additional features are speaker features.
- **Context Layer:** Apply 2-stack BI-LSTM to learn context representation
- **Output Layer:** Output nugget distribution for all utterances by softmax

Experiments

Table 3 shows the performance of ND subtask in JSD and RNSS

- **BL-BERT:** Simple BERT without any context or memory layer
- **BERT-LSTM:** Replace the embedding layer and utterance layer of HCNN-LSTM with BERT

Table 4 shows that multi-stack mechanism improves JSD & RNSS but gating mechanism and memory layer drop the performance

Table 3. Performance of ND subtask

Model	JSD	RNSS
BL-uniform	0.2304	0.3708
BL-popularity	0.1665	0.2653
BL-lstm	0.0248	0.0952
BL-BERT	0.0341	0.1171
HCNN-LSTM	0.0246	0.0962
BERT-LSTM	0.0228	0.0933

Table 4. Experiments of BERT-LSTM

Model	JSD	RNSS
BERT-LSTM	0.0228	0.0933
W/ gating	0.0244	0.0960
W/ memory layer	0.0234	0.0941
W/o multi-stack	0.0246	0.0951

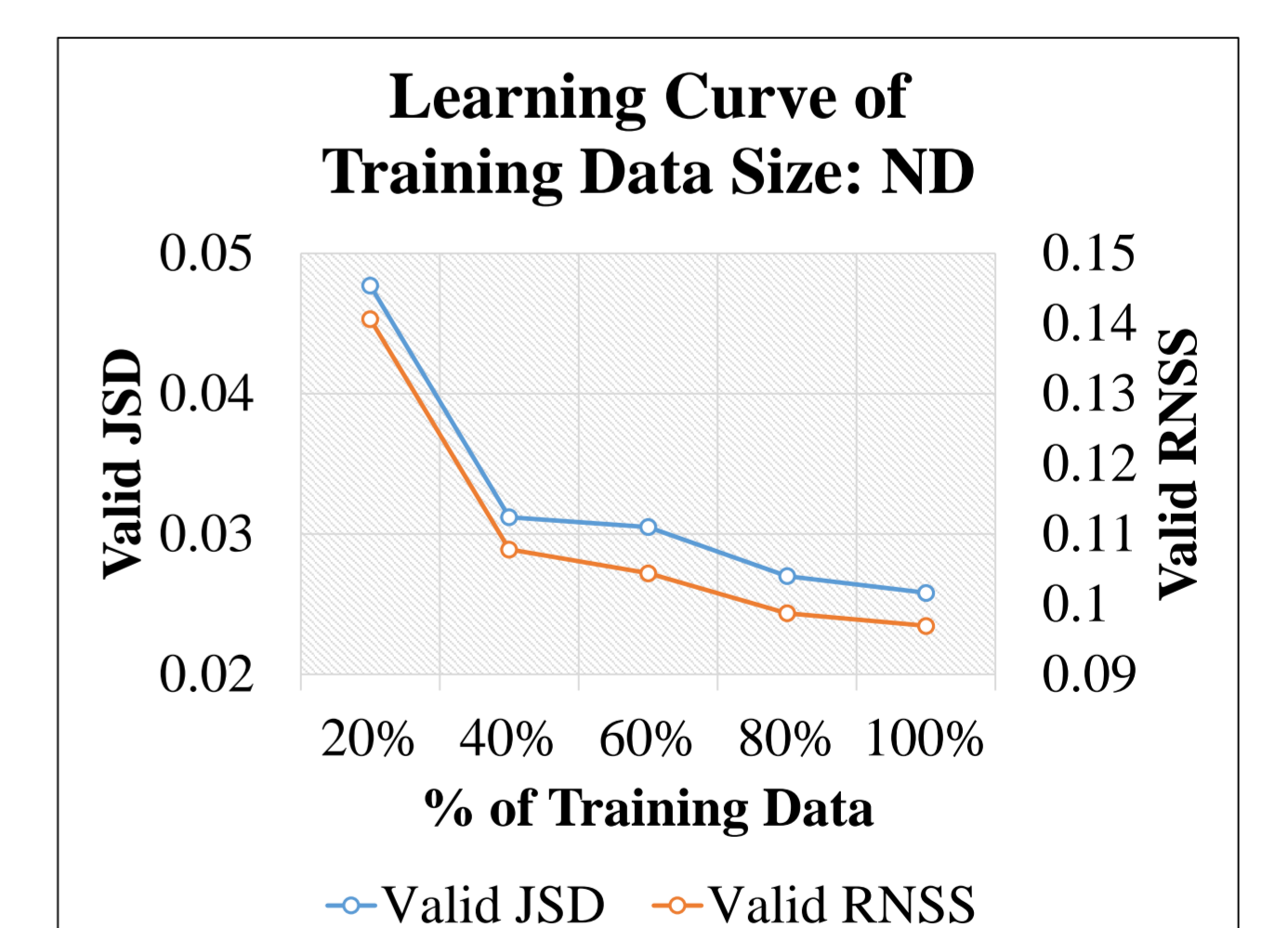


Fig 3. Learning Curve of BERT-LSTM

Conclusion

1. We propose two hierarchical models for ND and DQ subtasks
2. We compare models w/ & w/o gating mechanism & memory enhance
 - Both improve the performance of DQ subtask
 - But drop the performance of ND subtask
3. Data for ND might be insufficient which overfits in complex models
4. We compare sentence representation between BERT and HCNN
 - BERT as sentence representation performs better
5. Our models outperform other NTCIR baseline models in DQ & ND