NVIDIA AI CONFERENCE 2019

NVIDIA

Research in Natural Language Processing

Seunghyun Yoon

Machine Intelligence Lab.

SEOUL NATIONAL UNIVERSITY

Index

- Trend in NLP
- Question Answering System

- Extends NLP to other Area
- Multimodal Speech Emotion Recognition
- Conclusion

ME

David Seunghyun Yoon

PhD Student '17-

- Question Answering System
  - Answer-Selection QA
  - Machine Reading QA
- Multimodal Speech Emotion Recognition

Senior Engineer '06-'17

- Samsung Research (AI Team)
  - Question Answering System
  - Social Service (FE/BE)

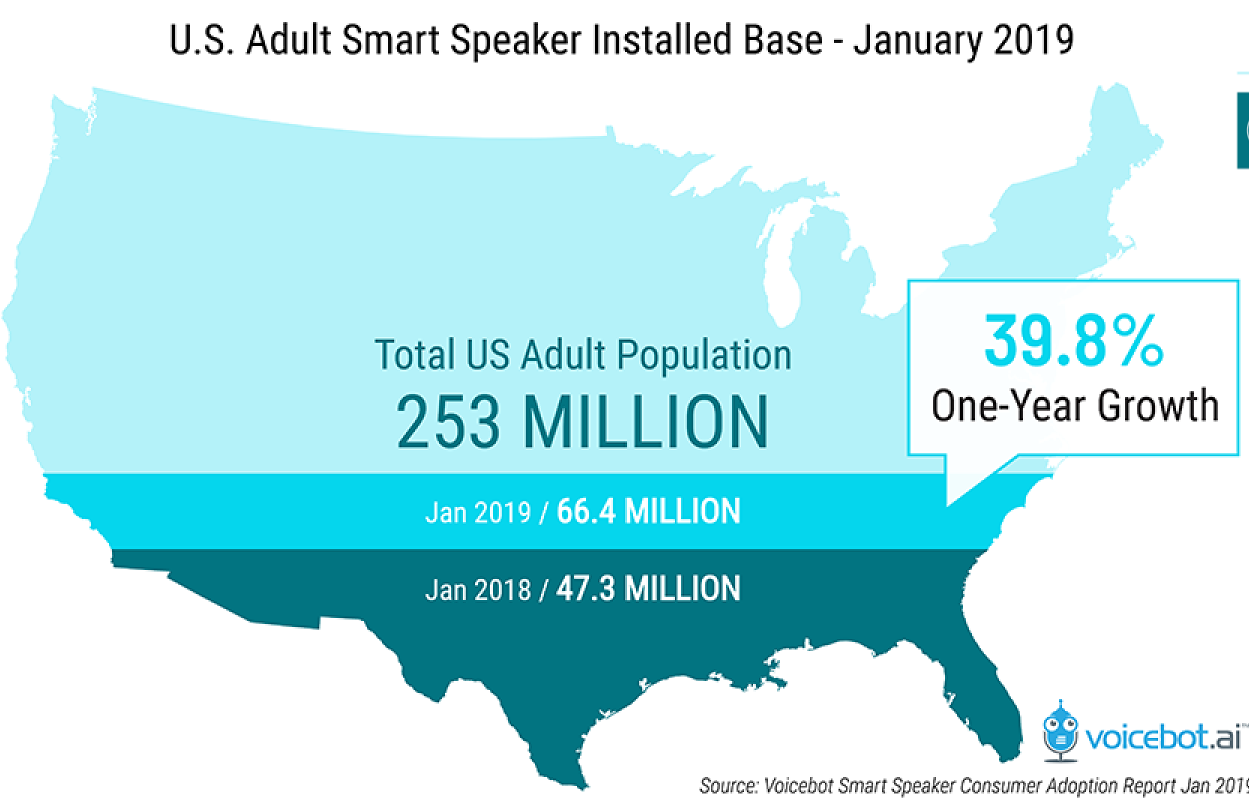
http://david-yoon.github.io/

The Era of AI

- Virtual Assistants are familiar to the customer
- Natural language I/F, Question Answering

The Era of AI

- **Virtual Assistants are familiar to the customer**
- **Smart speaker sales vaulted ownership to 26.2%**

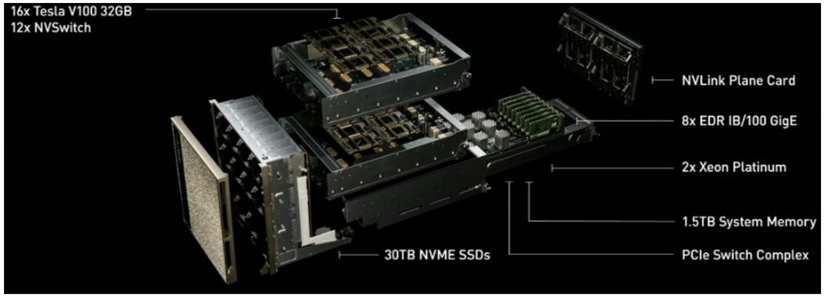


https://voicebot.ai/2019/03/07/u-s-smart-speaker-ownership-rises-40-in-2018-to-66-4-million-and-amazon-echo-maintains-market-share-lead-says-new-report-from-voicebot/

Deep Learning

- **Massive Computing Power**

Nvidia Tesla v100



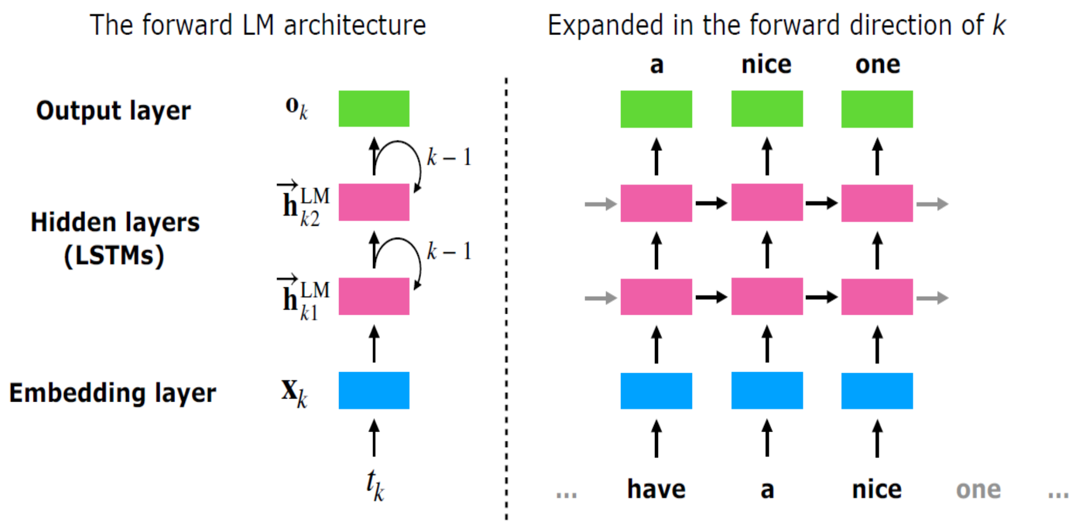
Nvidia DGX-2 (16 GPUs)

- **Large Dataset**

Outstanding Pre-trained Model

- **Deep contextualized word representation (ELMo), Peters et al., NAACL-18 Best Paper (Allen Institute, Univ. Washington)**

ELMo **(93.6 million parameters)**

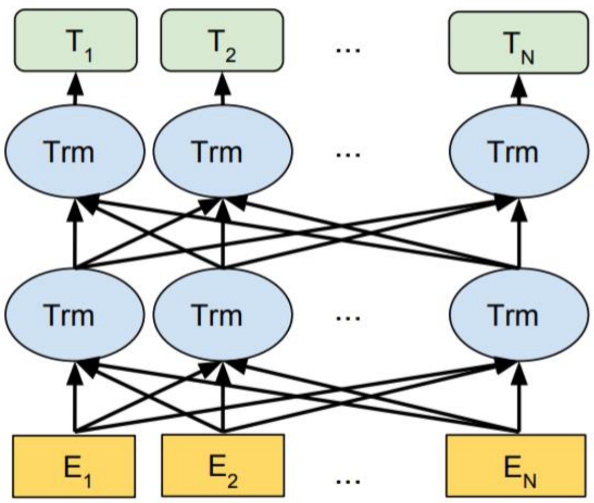


Peters, Matthew, et al. "Deep Contextualized Word Representations." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* 2018.

Outstanding Pre-trained Model

- **Bert: Pre-training of deep bidirectional transformers for language understanding, Devlin et al., NAACL-19 Best Paper** (google AI language)

BERT



Parameters:
- **340 million parameters**

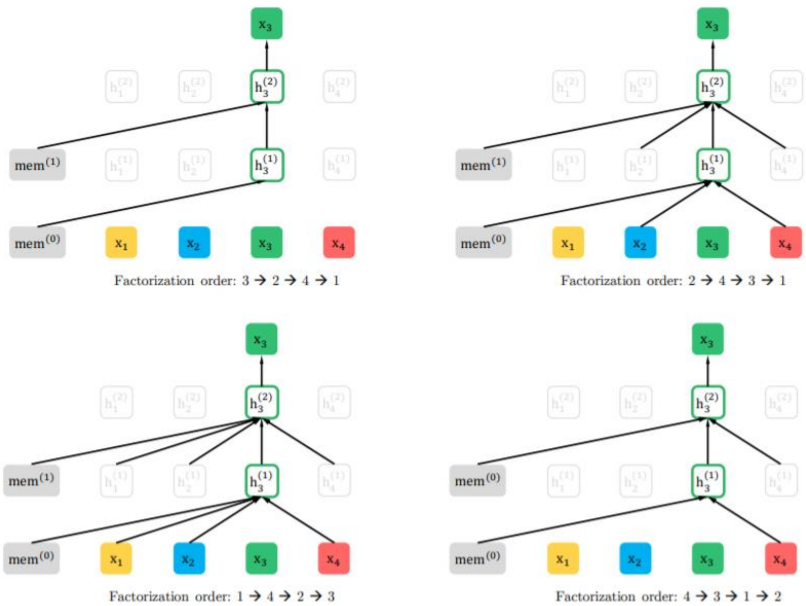
Training:
- **64 TPU chips**
- **4 days**

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Outstanding Pre-trained Model

- **XLNet: Generalized Autoregressive Pretraining for Language Understanding, Yang et al., Arxiv 19-06-19 (CMU, Google Brain)**

XLNet



Parameters:
- 340 million parameters

Training:
- 512 TPU v3 chips for 500K steps
- 2.5 days

512 TPU * 2.5 days * $8 a TPU
= $245,000

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Question Answering System (QA)

is a computer science discipline

within the fields of information retrieval (IR) and natural language processing (NLP),

which is concerned with building systems that automatically answer questions posed by humans in a natural language*.

*https://en.wikipedia.org/wiki/Question_answering

Two Major Research Direction in Academia

① Machine Reading QA

② Information retrieval (IR)-based QA

Two Major Research Direction in Academia

① Machine Reading QA

② Information retrieval (IR)-based QA

① Machine Reading QA

Given Passage, Question → Find the answer (fine-grained)

Passage: Tesla later approached Morgan to ask for more funds to build a more powerful transmitter. When asked where all the money had gone, *Tesla responded by saying that he was affected by the Panic of 1901, which he (Morgan) had caused.* Morgan was shocked by the reminder of his part in the stock market crash and by Tesla's breach of contract by asking for more funds. Tesla wrote another plea to Morgan, but it was also fruitless. Morgan still owed Tesla money on the original agreement, and Tesla had been facing foreclosure even before construction of the tower began.

Question: On what did Tesla blame for the loss of the initial money?

① Machine Reading QA

Given Passage, Question → Find the answer (fine-grained)

Passage: Tesla later approached Morgan to ask for more funds to build a more powerful transmitter. When asked where all the money had gone, *Tesla responded by saying that he was affected by the* ***Panic of 1901****, which he (Morgan) had caused.* Morgan was shocked by the reminder of his part in the stock market crash and by Tesla's breach of contract by asking for more funds. Tesla wrote another plea to Morgan, but it was also fruitless. Morgan still owed Tesla money on the original agreement, and Tesla had been facing foreclosure even before construction of the tower began.

Question: On what did Tesla blame for the loss of the initial money?

Answer: Panic of 1901

① Machine Reading QA

Q: Do we have a well-studied model?

A: Yes (for some dataset)

All year around competition

- SQuAD 1.0 / 2.0 (Stanford, 100K)
- MS-MARCO (MS, 1M)

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
4 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
5 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035
5 Jan 14, 2019	BERT + MMFT + ADA (single model) *Microsoft Research Asia*	83.040	85.892

SQuAD dataset leaderboard
https://rajpurkar.github.io/SQuAD-explorer/

① Machine Reading QA

Q: Can we apply it to the product?

A: Not yet (for some reasons)

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarter-back ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Br[illegible] Bowl XXXIII at age 38 and is c[illegible] tive Vice President of Football [illegible] Manager. Quarterback Jeff De[illegible] in Champ Bowl XXXIV.*"
Question: "*What is the name[illegible] was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Model	Original	ADDSENT	ADDONESENT
ReasoNet-E	**81.1**	39.4	49.8
SEDT-E	80.1	35.0	46.5
BiDAF-E	80.0	34.2	46.9
Mnemonic-E	79.1	**46.2**	**55.3**
Ruminating	78.8	37.4	47.7
[illegible]	[illegible]	37.9	47.0
[illegible]	[illegible]	**46.6**	**56.0**
[illegible]	[illegible]	39.4	50.3
[illegible]	[illegible]	40.3	50.0
[illegible]	[illegible]	33.9	44.8
[illegible]	[illegible]	39.5	49.5
[illegible]	[illegible]	34.3	45.7
Match-E	[illegible]	[illegible]	41.8
Match-S	[illegible]	[illegible]	39.0
DCR	69.3	37.8	45.1
Logistic	50.4	23.2	30.4

Pretrained model is fooled by the addition of an adversarial distracting sentence

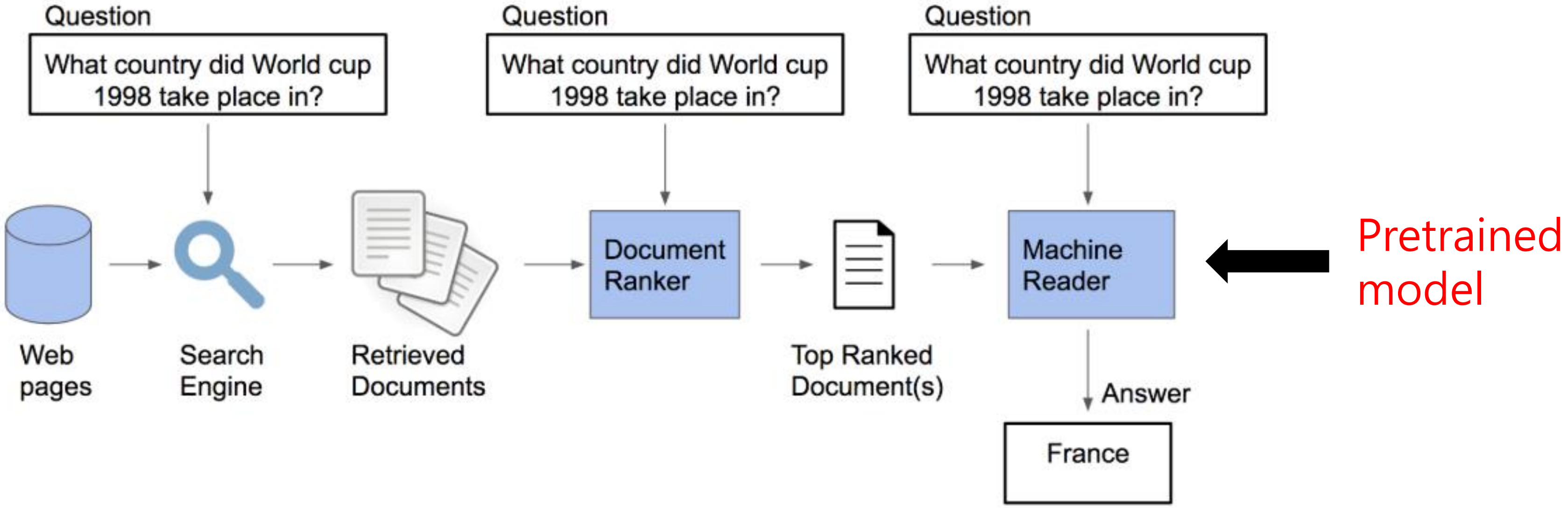
degradation

ref: Adversarial Examples for Evaluating Reading Comprehension Systems, Jia et. al., EMNLP-17 outstanding paper

① Machine Reading QA

Q: Can we apply it to the product?

A: Not yet (for some reasons)



ref: Training a Ranking Function for Open-Domain Question Answering, Htut. et. al., NAACL-18

① Machine Reading QA

Q: Can we apply it to the product?

A: Not yet (for some reasons)

SQuAD dataset results

	EM	F1
BiDAF (Seo 2016)	68.0	77.5

Open domain results

	EM	F1
GA (Dhingra et al., 2017a)	26.4	26.4
BiDAF (Seo et al., 2016)	25.9	28.5
R^3 (Wang et al., 2017)	**35.3**	**41.7**
SR^2 (Wang et al., 2017)	31.9	38.7

ref: Training a Ranking Function for Open-Domain Question Answering, Htut. et. al., NAACL-18

Two Major **Research Direction** in Academia

① Machine Reading QA

② Information retrieval (IR)-based QA

② IR-based QA

Given Passage, Question → Find the answer (coarse-level)

Passage: Journey to the West is one of the four classics of Chinese literature. Written by the Ming Dynasty novelist Wu Cheng'en during the 16th century, this beloved adventure tale combines action, humor, and spiritual lessons.

The novel takes place in the seventh century. It tells the story of one of Buddha Sakyamuni's disciples who was banished from the heavenly paradise for the crime of slighting the Buddha Law. He was sent to the human world and forced to spend ten lifetimes practicing religious self-cultivation in order to atone for his sins.

In his tenth lifetime, now during the Tang Dynasty, he reincarnates as a monk named Xuan Zang (also known as Tang Monk and Tripitaka). The emperor wishes this monk can travel west and bring holy Mahayana Buddhist scriptures back to China. After being inspired by a vision from the Bodhisattva Guanyin, the monk accepts the mission and sets off on the sacred quest.

Question: Who is the Tang?

② IR-based QA

Split the passage into multiple sentences → focus on the relevant one

Passage: Journey to the West is one of the four classics of Chinese literature. Written by the Ming Dynasty novelist Wu Cheng'en during the 16th century, this beloved adventure tale combines action, humor, and spiritual lessons.

The novel takes place in the seventh century. It tells the story of one of Buddha Sakyamuni's disciples who was banished from the heavenly paradise for the crime of slighting the Buddha Law. He was sent to the human world and forced to spend ten

sentence-level

cticing religious self-cultivation in order to atone for his sins.

In his tenth lifetime, now during the Tang Dynasty, he reincarnates as a monk named Xuan Zang (also known as Tang Monk and Tripitaka). The emperor wishes this monk can travel west and bring holy Mahayana Buddhist scriptures back to China. After being inspired by a vision from the Bodhisattva Guanyin, the monk accepts the mission and sets off on the sacred quest.

Question: Who is the Tang?

② IR-based QA

Model has more information to consider (paragraph > sentence)

Passage: Journey to the West is one of the four classics of Chinese literature. Written by the Ming Dynasty novelist Wu Cheng'en during the 16th century, this beloved adventure tale combines action, humor, and spiritual lessons.

The novel takes place in the seventh century. It tells the story of one of Buddha Sakyamuni's disciples who was banished from the heavenly paradise for the crime of slighting the Buddha Law. He was sent to the human world and forced to spend ten lifetimes practicing religious self-cultivation in order to atone for his s

paragraph-level

In his tenth lifetime, now during the Tang Dynasty, he reincarnates as a monk named Xuan Zang (also known as Tang Monk and Tripitaka). The emperor wishes this monk can travel west and bring holy Mahayana Buddhist scriptures back to China. After being inspired by a vision from the Bodhisattva Guanyin, the monk accepts the mission and sets off on the sacred quest.

Question: Who is the Tang?

② IR-based QA

Q: Is well-studied model available?

A: Yes (for sentence-level)

Long-history dataset

- TREC-QA since 04′ (1.2K)
- WikiQA since 15′ (1k)

Algorithm - Clean Version of TREC QA	Reference	MAP	MRR
W&I (2015)	Wang and Ittycheriah (2015)	0.746	0.820
Tan (2015) - QA-LSTM/CNN+attention	Tan et al. (2015)	0.728	0.832
dos Santos (2016) - Attentive Pooling CNN	dos Santos et al. (2016)	0.753	0.851
Wang et al. (2016) - L.D.C Model	Wang et al. (2016)	0.771	0.845
H&L (2015) - Multi-Perspective CNN	He and Lin (2015)	0.777	0.836
Tay et al. (2017) - HyperQA (Hyperbolic Embeddings)	Tay et al. (2017)	0.784	0.865
Rao et al. (2016) - PairwiseRank + Multi-Perspective CNN	Rao et al. (2016)	0.801	0.877
Wang et al. (2017) - BiMPM	Wang et al. (2017)	0.802	0.875
Bian et al. (2017) - Compare-Aggregate	Bian et al. (2017)	0.821	0.899
Shen et al. (2017) - IWAN	Shen et al. (2017)	0.822	0.889
Tran et al. (2018) - IWAN + sCARNN	Tran et al. (2018)	0.829	0.875
Tay et al. (2018) - Multi-Cast Attention Networks (MCAN)	Tay et al. (2018)	0.838	0.904
Tayyar Madabushi (2018) - Question Classification + PairwiseRank + Multi-Perspective CNN	Tayyar Madabushi et al. (2018)	0.865	0.904
Yoon et al. (2019) - Compare-Aggregate + LanguageModel + LatentClustering	Yoon et al. (2019)	0.868	0.928

https://aclweb.org/aclwiki/Question_Answering_(State_of_the_art)

Research Objective?

- **Consider the Answer Span**

model complexity ↔ robustness

Passage: Tesla later approached Morgan to ask for more funds to build a more powerful transmitter. When asked where all the money had gone, *Tesla responded by saying that he was affected by the Panic of 1901, which he (Morgan) had caused*. Morgan was shocked by the reminder of his part in the stock market crash and by Tesla's breach of contract by asking for more funds. Tesla wrote another plea to Morgan, but it was also fruitless. Morgan still owed Tesla money on the original agreement, and Tesla had been facing foreclosure even before construction of the tower began.

exact answer

Passage: Journey to the West is one of the four classics of Chinese literature. Written by the Ming Dynasty novelist Wu Cheng'en during the 16th century, this beloved adventure tale combines action, humor, and spiritual lessons.

In his tenth lifetime, now during the Tang Dynasty, he reincarnates as a monk named Xuan Zang (also known as Tang Monk and Tripitaka). The emperor wishes this monk can travel west and bring holy Mahayana Buddhist scriptures back to China. After being inspired by a vision from the Bodhisattva Guanyin, the monk accepts the mission and sets off on the sacred quest.

sentence-level

Passage: Journey to the West is one of the four classics of Chinese literature. Written by the Ming Dynasty novelist Wu Cheng'en during the 16th century, this beloved adventure tale combines action, humor, and spiritual lessons.

In his tenth lifetime, now during the Tang Dynasty, he reincarnates as a monk named Xuan Zang (also known as Tang Monk and Tripitaka). The emperor wishes this monk can travel west and bring holy Mahayana Buddhist scriptures back to China. After being inspired by a vision from the Bodhisattva Guanyin, the monk accepts the mission and sets off on the sacred quest.

paragraph-level

Who Leads the NLP Tasks?

- **Power of Pre-trained Model (Machine Reading QA)**

Squad 2.0 Leaderboard*

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 May 21, 2019	XLNet (single model) *XLNet Team*	86.346	89.133
5 Apr 13, 2019	SemBERT(ensemble) *Shanghai Jiao Tong University*	86.166	88.886
5 May 14, 2019	SG-Net (ensemble) *Anonymous*	86.211	88.848
6 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621
7 May 14, 2019	SG-Net (single model) *Anonymous*	85.229	87.926
8	SemBERT (single model)	84.800	87.864

*https://rajpurkar.github.io/SQuAD-explorer/

Who Leads the NLP Tasks?

- **Power of Pre-trained Model (Various NLP Tasks)**

GLUE Leaderboard*

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI
	1	XLNet Team	XLNet-Large (ensemble)		88.4	67.8	96.8	93.0/90.7	91.6/91.1	74.2/90.3	90.2	89.8	98.6	86.3	90.4
+	2	Microsoft D365 AI & MSR	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0
	3	GLUE Human Baselines	GLUE Human Baselines		87.1	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
+	4	王玮	ALICE large ensemble (Alibaba D		86.3	68.6	95.2	92.6/90.2	91.1/90.6	74.4/90.7	88.2	87.9	95.7	83.5	80.8
	5	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1
	6	张倬胜	SemBERT		82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1
	7	Anonymous Anonymous	BERT + BAM		82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1

*https://gluebenchmark.com/leaderboard

Our Results on IR-based QA

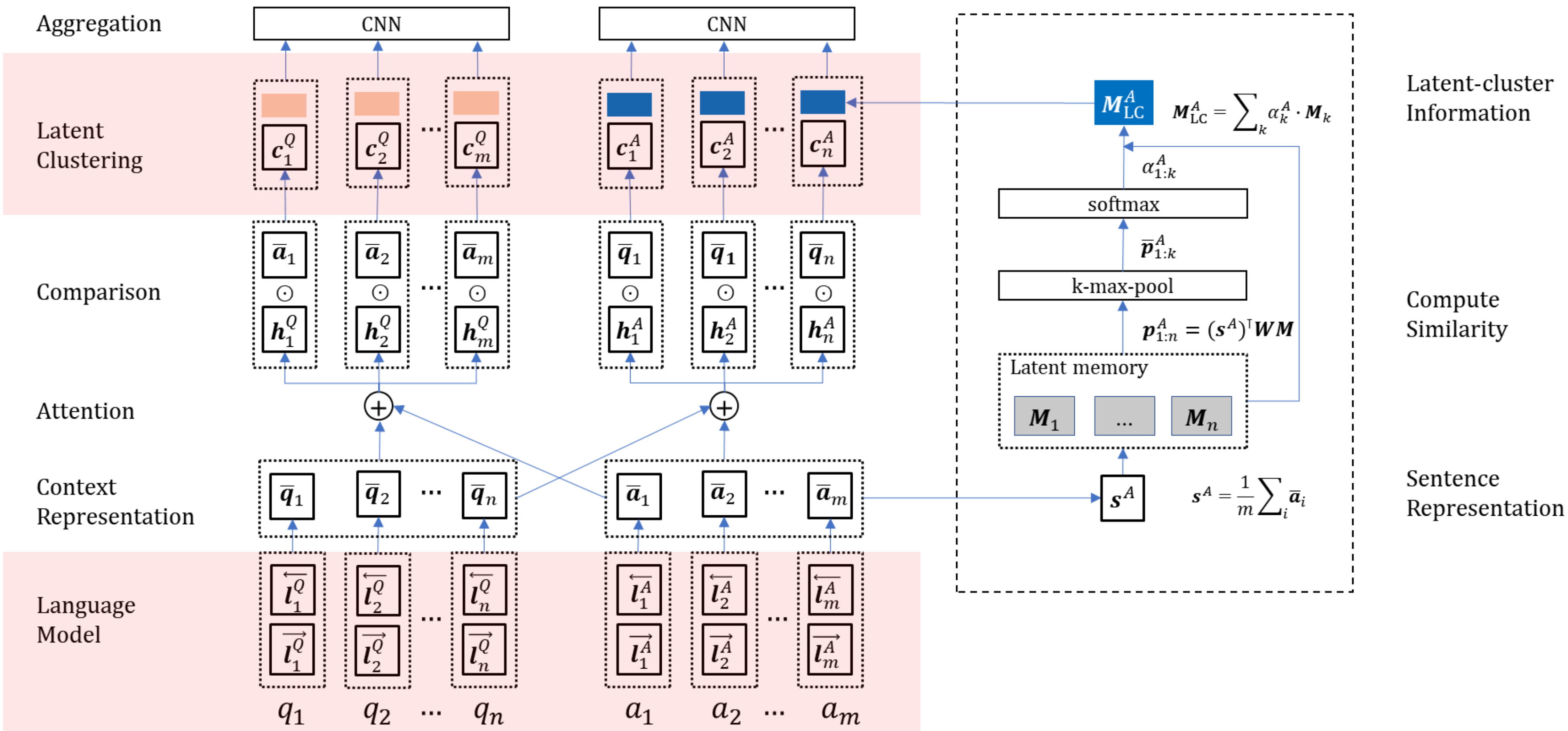
- **Power of Pre-trained Model (Answer-selection QA)**

- Yoon, et al. "**A Compare-Aggregate Model with Latent Clustering for Answer Selection**." *arXiv preprint arXiv:1905.12897* (2019).

- **Main Ideas**
  - Adopt the pre-trained Language Model (**LM**)
  - Apply Transfer-Learning (**TL**) using QNLI dataset
  - Apply Latent Cluster method (**LC**)

Model for the Answer-Selection QA

- **Power of Pre-trained Model (Answer-selection QA)**



Experimental Results

- **We achieve the state-of-the-art performance in both dataset**

Model	Wiki QA				TREC-QA			
	MAP		MRR		MAP		MRR	
	dev	test	dev	test	dev	test	dev	test
Compare-Aggregate (2016) [1]	0.743		0.754		-		-	
Comp-Clip (2017) [2]	0.754		0.764		0.821		0.899	
IWAN (2017) [3]	0.733		0.750		0.822		0.899	
IWAN + sCARNN (2018) [4]	0.716*		0.722*		0.829		0.875	
MCAN (2018) [5]	-		-		0.838		0.904	
Question Classification (2018) [6]	-		-		0.865		0.904	
List-wise Learning to Rank								
Comp-Clip (our implementation)	0.756	0.708	0.766	0.725	0.750	0.744	0.805	0.791
Comp-Clip (our implementation) + LM	0.783	0.748	0.791	0.768	[illegible]	0.823	0.870	0.868
Comp-Clip (our implementation) + LM + LC	0.787	**0.759**	0.793	**0.772**	[illegible]	[illegible]	[illegible]	[illegible]
Comp-Clip (our implementation) + LM + LC +TL	0.820	**0.825**	0.826	**0.837**	[illegible]	0.848	0.911	0.902
Point-wise Learning to Rank								
Comp-Clip (our implementation)	0.776	0.714	0.784	0.732	[illegible]	0.835	0.933	0.877
Comp-Clip (our implementation) + LM	0.785	0.746	0.789	0.762	0.872	0.850	[illegible]	[illegible]
Comp-Clip (our implementation) + LM + LC	0.794	0.754	0.798	0.771	0.883	**0.858**	0.955	**0.923**
Comp-Clip (our implementation) + LM + LC +TL	0.827	0.814	0.828	0.827	0.906	**0.874**	0.974	**0.929**

Language model
+ topic model

7.2% (0.708 → 0.759)

8.6% (0.759 → 0.825)

Additional dataset

LM: Language Model
LC : Latent Clustering
TL : Transfer Learning (using Squad-T)

Experimental Results

- **We achieve the state-of-the-art performance in both dataset**

Model	Wiki QA				TREC-QA			
	MAP		MRR		MAP		MRR	
	dev	test	dev	test	dev	test	dev	test
Compare-Aggregate (2016) [1]		0.743		0.754		-		-
Comp-Clip (2017) [2]		0.754		0.764		0.821		0.899
IWAN (2017) [3]		0.733		0.750		0.822		0.899
IWAN + sCARNN (2018) [4]		0.716*		0.722*		0.829		0.875
MCAN (2018) [5]		-		-		0.838		0.904
Question Classification (2018) [6]		-		-		0.865		0.904
List-wise Learning to Rank								
Comp-Clip (our implementation)	0.756	0.708	0.766	0.725	0.750	0.744	0.805	0.791
Comp-Clip (our implementation) + LM	0.783	0.748	0.791	0.768	0.825	0.823	0.870	0.868
Comp-Clip (our implementation) + LM + LC	0.787	**0.759**	0.793	**0.772**	0.841	0.832	0.842	0.880
Comp-Clip (our implementation) + LM + LC +TL	0.820	**0.825**	0.826	**0.837**	0.866	0.848	0.911	0.902
Point-wise Learning to Rank								
Comp-Clip (our implementation)	[illegible]	[illegible]	[illegible]	0.732	0.866	0.835	0.933	0.877
Comp-Clip (our implementation) + LM	0.785	0.746	0.789	0.762	0.872	0.850	0.930	0.898
Comp-Clip (our implementation) + LM + LC	0.794	0.754	0.798	0.771	0.883	**0.858**	0.955	**0.923**
Comp-Clip (our implementation) + LM + LC +TL	[illegible]	0.814	0.828	0.827	0.906	**0.874**	0.974	**0.929**

Language model + topic model

2.7% (0.835 → 0.858)

1.8% (0.858 → 0.874)

Additional dataset

LM: Language Model
LC : Latent Clustering
TL : Transfer Learning (using Squad-T)

Specific Task can be tackled via

Model (Researchers)

Data (Service)

Implementation (Engineers)

Computing Resources

Extends NLP to other Area

Speech Emotion Recognition

Exploiting textual and acoustic data of an utterance for the speech emotion classification task

Extends NLP to other Area

Speech Emotion Recognition Using Multi-hop Attention Mechanism

ICASSP-2019

[1]Seunghyun Yoon, **[1]Seokhyun Byun**, **[2]Subhadeep Dey** and **[1]Kyomin Jung**

SEOUL NATIONAL UNIVERSITY

idiap RESEARCH INSTITUTE

Research Problem

Speech Emotion Recognition

Exploiting textual and acoustic data of an utterance for the speech emotion classification task

Dataset

- **Interactive Emotional Dyadic Motion Capture (IEMOCAP)**
  - **Five sessions** of utterances between two speakers (one male and one female)
  - Total 10 unique speakers participated

- **Environment setting**
  - **1,636 happy, 1,084 sad, 1,103 angry and 1,708 neutral**
  - “**excitement**” → merge with “**happiness**”
  - **10-fold** cross-validation

Related Work: Single modality

- **Using Regional Saliency for Speech Emotion Recognition**, Aldeneh, et., al., ICASSP-17

- **CNN based model**
- Achieve up to **60.7%** WA in IEMOCAP dataset

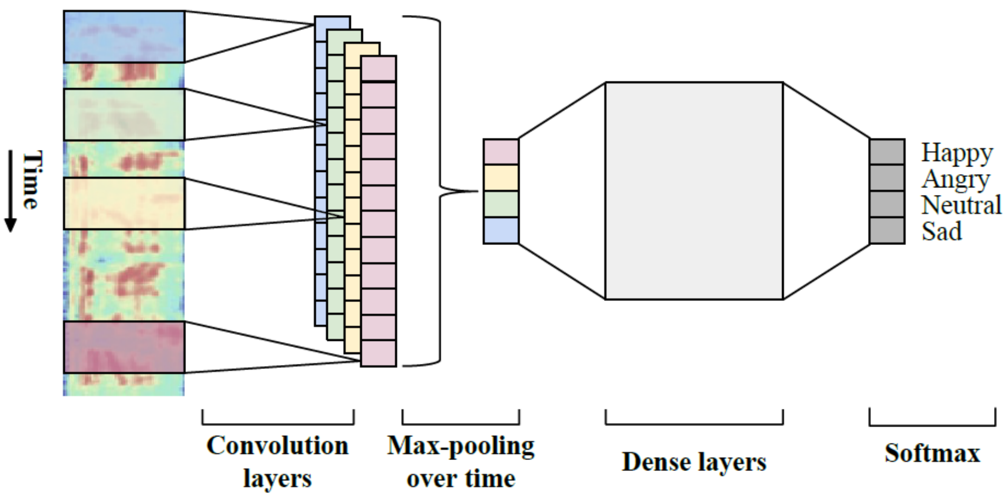
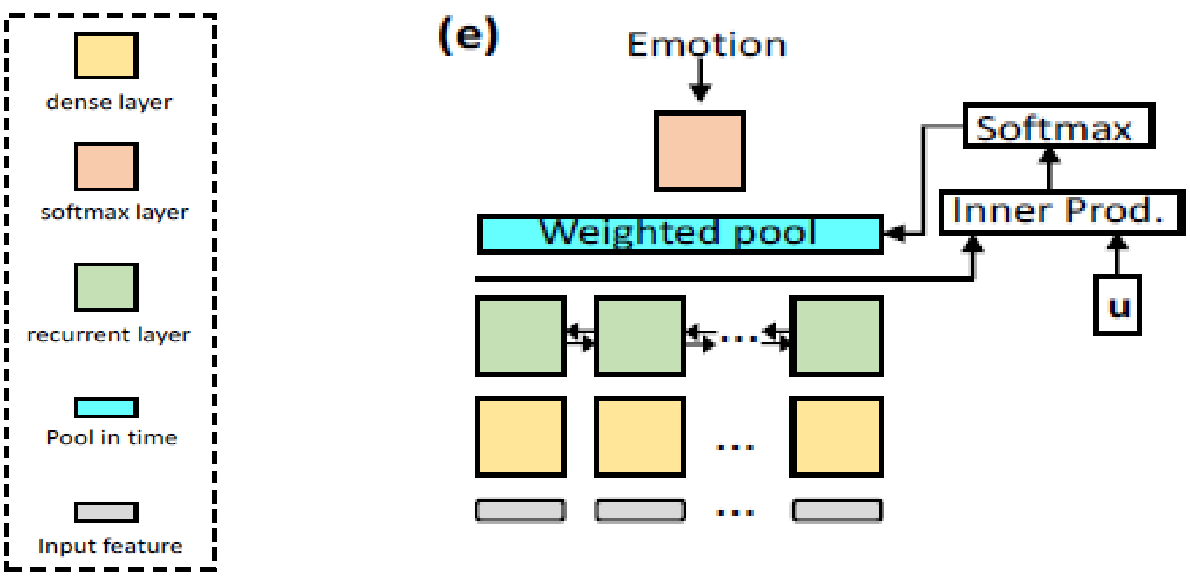


Fig. 1. Network Overview (four filters shown).

Related Work: Single modality

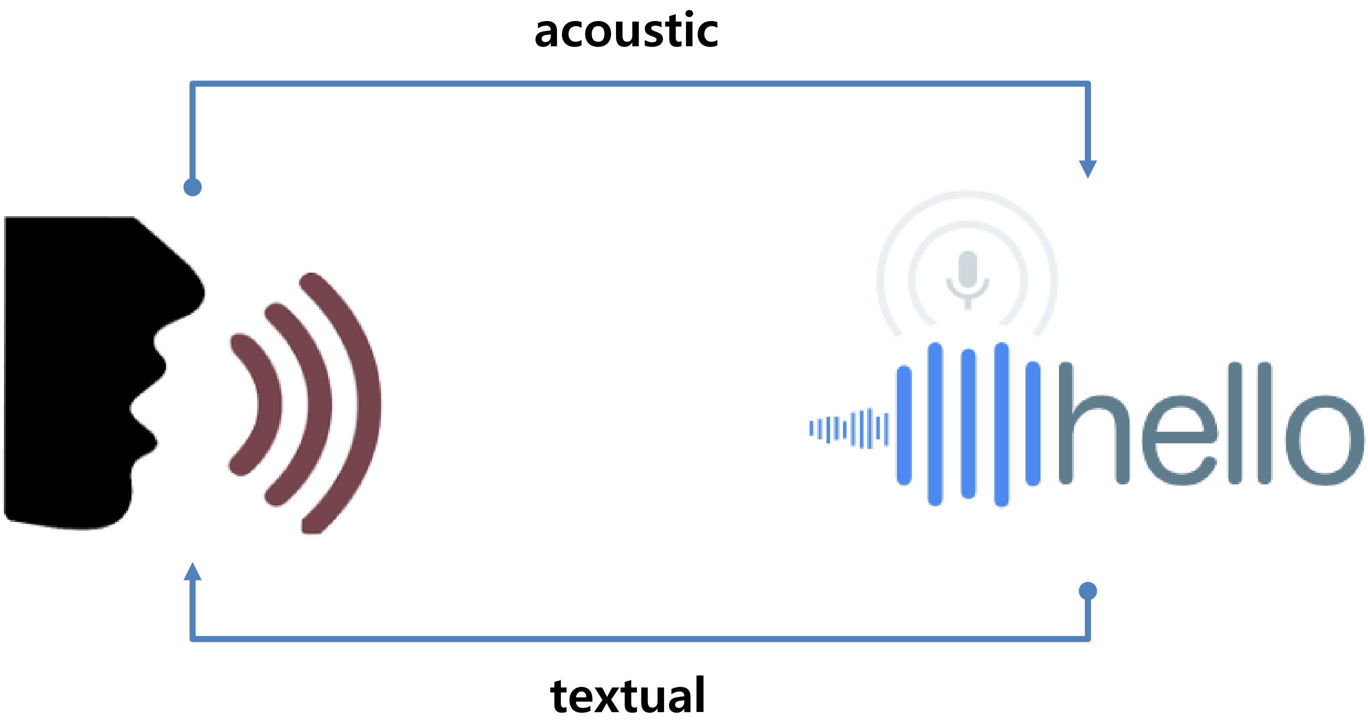
- **Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention**, Mirsamadi et., al., ICASSP-17

- **RNN based model with Attention mechanism**
- Achieve up to **63.5%** WA in IEMOCAP dataset



Our Idea

- Motivated by human behavior
  - Contextual Understanding from an iterative process



Bidirectional Recurrent Encoder (BRE)

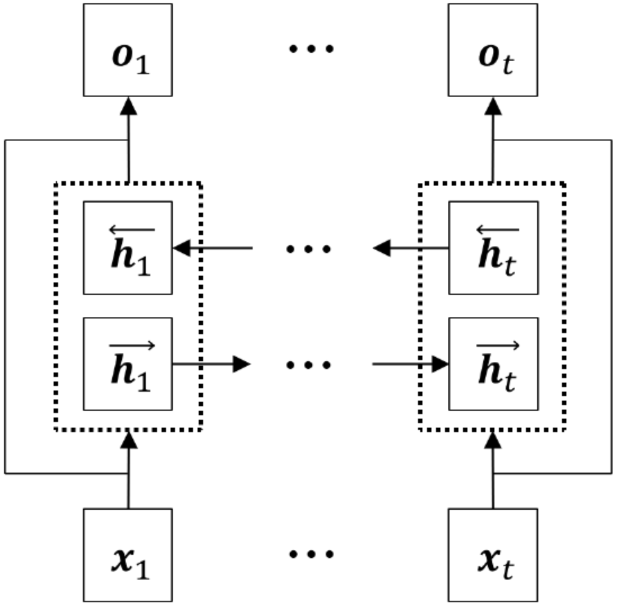
- Audio-BRE
  - Recurrent Encoder for audio modality
  - Bidirectional
  - Residual Connection

$$\overrightarrow{\mathbf{h}}_t = f_\theta(\overrightarrow{\mathbf{h}}_{t-1}, \overrightarrow{\mathbf{x}}_t) + \overrightarrow{\mathbf{x}}_t,$$

$$\overleftarrow{\mathbf{h}}_t = f'_\theta(\overleftarrow{\mathbf{h}}_{t+1}, \overleftarrow{\mathbf{x}}_t) + \overleftarrow{\mathbf{x}}_t,$$

$$\mathbf{o}_t = [\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t],$$

$$\mathbf{o}_t^A = [\mathbf{o}_t; \mathbf{p}]$$



BRE model

- Features

$\boldsymbol{x_t}$: audio feature (MFCC)
$\mathbf{p}$: prosodic feature vector

Bidirectional Recurrent Encoder (BRE)

- **Text-BRE**
  - Recurrent Encoder for textual modality

- **Tokenize textual information**
  - I'm happy to hear the story

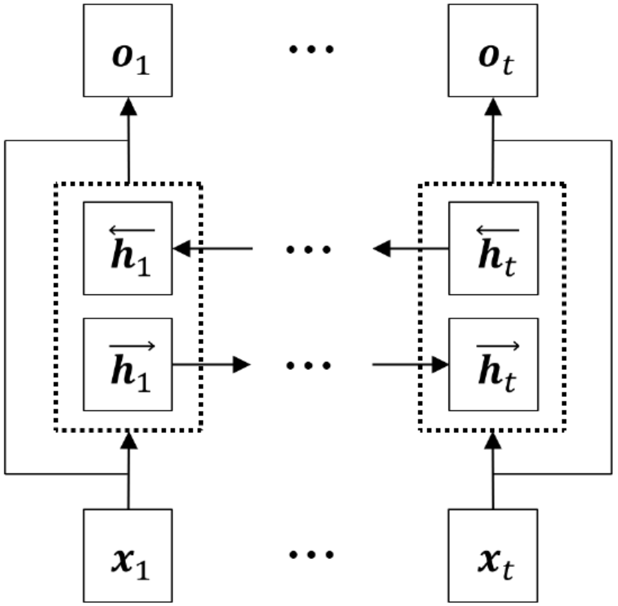
  → I 'm happy to hear the story

$$\overrightarrow{\mathbf{h}}_t = f_\theta(\overrightarrow{\mathbf{h}}_{t-1}, \overrightarrow{\mathbf{x}}_t) + \overrightarrow{\mathbf{x}}_t,$$

$$\overleftarrow{\mathbf{h}}_t = f'_\theta(\overleftarrow{\mathbf{h}}_{t+1}, \overleftarrow{\mathbf{x}}_t) + \overleftarrow{\mathbf{x}}_t,$$

$$\mathbf{o}_t^T = [\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t]$$

x_t : textual feature



BRE model

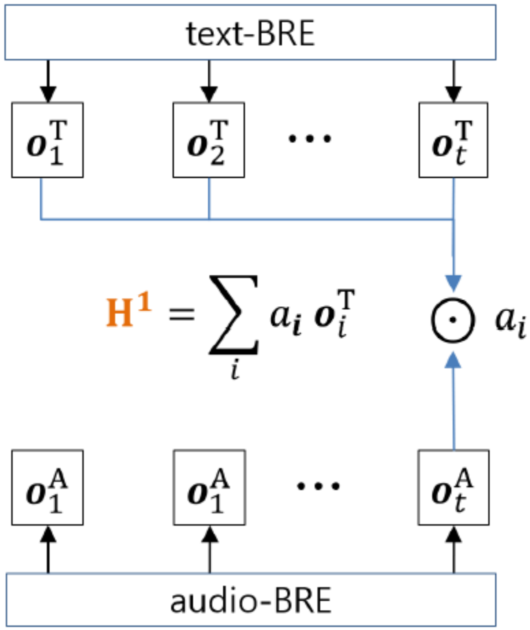
① Multi-hop Attention (MHA)

- **First Hop**

- **Context : Audio information**
- **Aggregate : Textual information**
- **Result : $\mathbf{H}^1$**

$$a_i = \frac{\exp(\ (\mathbf{o}^A_{\text{last}})^\intercal\ \mathbf{o}^T_i\ )}{\sum_i \exp(\ (\mathbf{o}^A_{\text{last}})^\intercal\ \mathbf{o}^T_i\ )},\ (i = 1, ..., t)$$

$$\mathbf{H}^1 = \sum_i a_i\ \mathbf{o}^T_i,\ \ \mathbf{H} = [\mathbf{H}^1; \mathbf{o}^A_{\text{last}}].$$



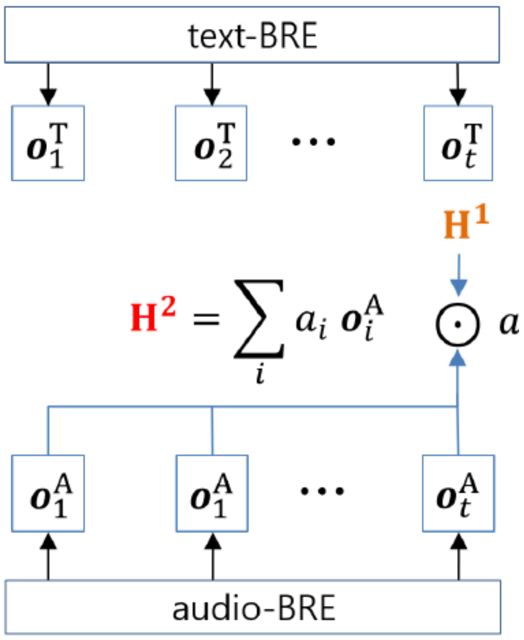
② Multi-hop Attention (MHA)

- **Second Hop**

- **Context : Updated textual information**
- **Aggregate : Audio information**
- **Result : $\mathbf{H}^2$**

$$a_i = \frac{\exp(\ (\mathbf{H}_1)^\intercal\ \mathbf{o}_i^A\ )}{\sum_i \exp(\ (\mathbf{H}_1)^\intercal\ \mathbf{o}_i^A\ )},\ (i = 1, ..., t)$$

$$\mathbf{H}^2 = \sum_i a_i\ \mathbf{o}_i^A,\ \ \mathbf{H} = [\mathbf{H}^1; \mathbf{H}^2],$$



Results

- **Textual information vs Acoustic information**
  - **text-BRE** shows higher performance than that of **audio-BRE** by 8%

Model	Modality	WA	UA
Ground-truth transcript			
E_vec-MCNN-LSTM [18]	A+T	0.649	0.659
MDRE [7]	A+T	0.718	-
audio-BRE (ours)	A	0.646	0.652
text-BRE (ours)	T	0.698	0.703
MHA-1 (ours)	A+T	**0.756**	**0.765**
MHA-2 (ours)	A+T	**0.765**	**0.776**
MHA-3 (ours)	A+T	0.740	0.753
ASR-processed transcript			
text-BRE-ASR (ours)	T	0.652	0.658
MHA-2-ASR (ours)	A+T	0.730	0.739

8% (0.646 → 0.698)

Results

- **Comparison with best baseline model**
  - **MHA-2** outperformed the **MDRE*** by 6.5%

Model	Modality	WA	UA
Ground-truth transcript			
E_vec-MCNN-LSTM [18]	A+T	0.649	0.659
MDRE [7]	A+T	0.718	-
audio-BRE (ours)	A	0.646	0.652
text-BRE (ours)	T	0.698	0.703
MHA-1 (ours)	A+T	**0.756**	**0.765**
MHA-2 (ours)	A+T	**0.765**	**0.776**
MHA-3 (ours)	A+T	0.740	0.753
ASR-processed transcript			
text-BRE-ASR (ours)	T	0.652	0.658
MHA-2-ASR (ours)	A+T	0.730	0.739

6.5% (0.718 → 0.765)

***MDRE (multimodal dual recurrent encoder), SLT-18 : previous state-of-the-art**

Results

- ASR-processed **transcript (WER 5.53%)**
  - performance degradation in **text-BRE-ASR** by 6.6%

Model	Modality	WA	UA
Ground-truth transcript			
E_vec-MCNN-LSTM [18]	A+T	0.649	0.659
MDRE [7]	A+T	0.718	-
audio-BRE (ours)	A	0.646	0.652
text-BRE (ours)	T	0.698	0.703
MHA-1 (ours)	A+T	**0.756**	**0.765**
MHA-2 (ours)	A+T	**0.765**	**0.776**
MHA-3 (ours)	A+T	0.740	0.753
ASR-processed transcript			
text-BRE-ASR (ours)	T	0.652	0.658
MHA-2-ASR (ours)	A+T	0.730	0.739

6.6% (0.698 → 0.652)

Results

- **ASR-processed transcript (WER 5.53%)**
  - performance degradation in **MHA-2-ASR** by 4.6%

Model	Modality	WA	UA
Ground-truth transcript			
E_vec-MCNN-LSTM [18]	A+T	0.649	0.659
MDRE [7]	A+T	0.718	-
audio-BRE (ours)	A	0.646	0.652
text-BRE (ours)	T	0.698	0.703
MHA-1 (ours)	A+T	**0.756**	**0.765**
MHA-2 (ours)	A+T	**0.765**	**0.776**
MHA-3 (ours)	A+T	0.740	0.753
ASR-processed transcript			
text-BRE-ASR (ours)	T	0.652	0.658
MHA-2-ASR (ours)	A+T	0.730	0.739

4.6% (0.765 → 0.730)

Results

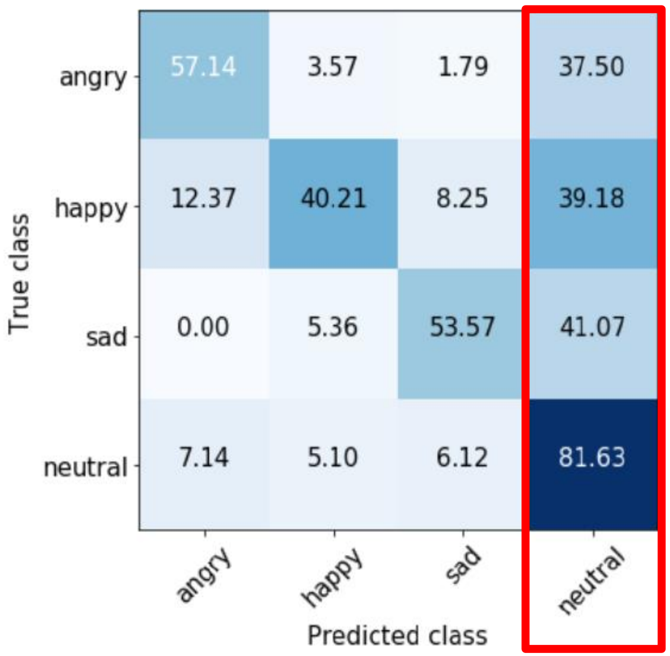
- **ASR-processed (WER 5.53%) vs ground-truth**
  - **MHA-2** still outperformed the **MDRE** by 1.6%

Model	Modality	WA	UA
Ground-truth transcript			
E_vec-MCNN-LSTM [18]	A+T	0.649	0.659
MDRE [7]	A+T	0.718	-
audio-BRE (ours)	A	0.646	0.652
text-BRE (ours)	T	0.698	0.703
MHA-1 (ours)	A+T	**0.756**	**0.765**
MHA-2 (ours)	A+T	**0.765**	**0.776**
MHA-3 (ours)	A+T	0.740	0.753
ASR-processed transcript			
text-BRE-ASR (ours)	T	0.652	0.658
MHA-2-ASR (ours)	A+T	0.730	0.739

1.6% **(0.718 → 0.730)**

Error Analysis

- **Audio-BRE**
  - Most of the emotion labels are frequently misclassified as ***"neutral"***
  - Supporting the claims in [7, 25]



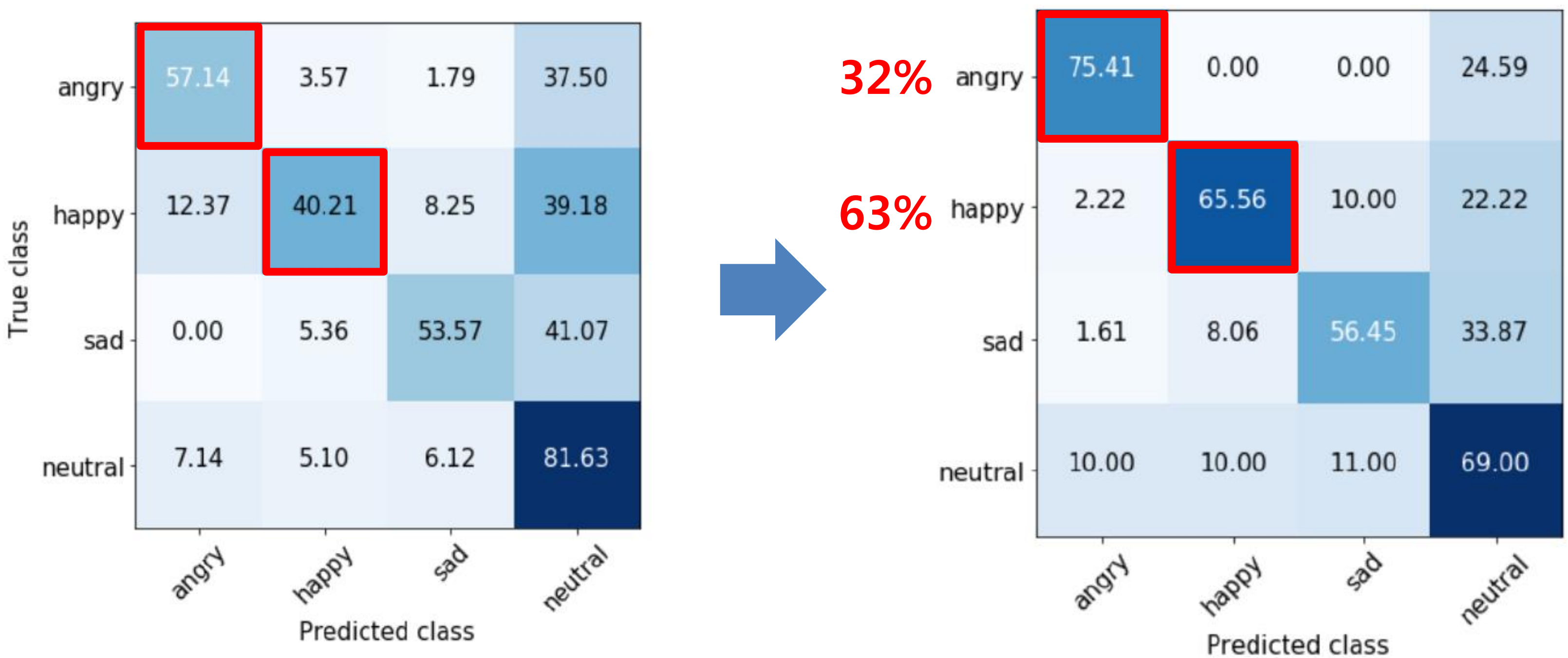
(a) audio-BRE

[7] Multimodal speech emotion recognition using audio and text, Yoon et. al., SLT-18

[25] Attentive convolutional
neural network based speech emotion recognition: A
study on the impact of input features, signal length, and acted speech, Neumann et. al., Interspeech-17

Error Analysis

- **Text-BRE**
  - *"**angry**"* and "*happy*" are correctly classified by 32% (57.14 to 75.41) and 63% (40.21 to 65.56)

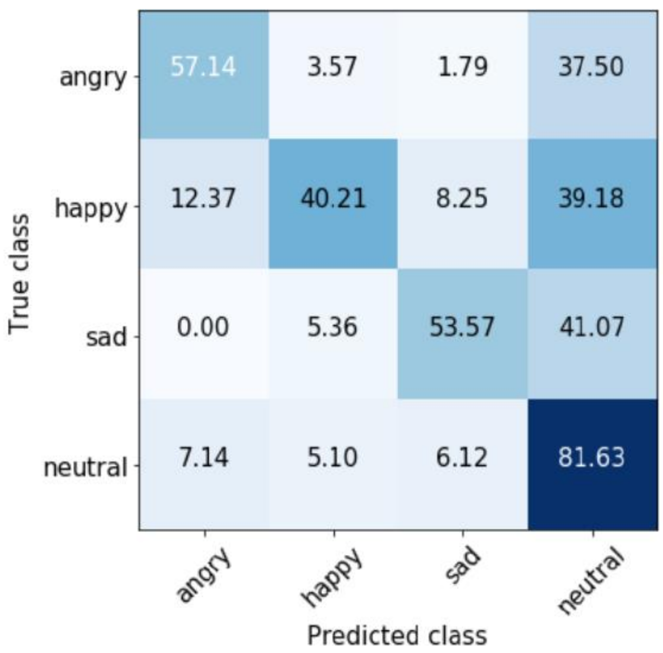


(a) audio-BRE

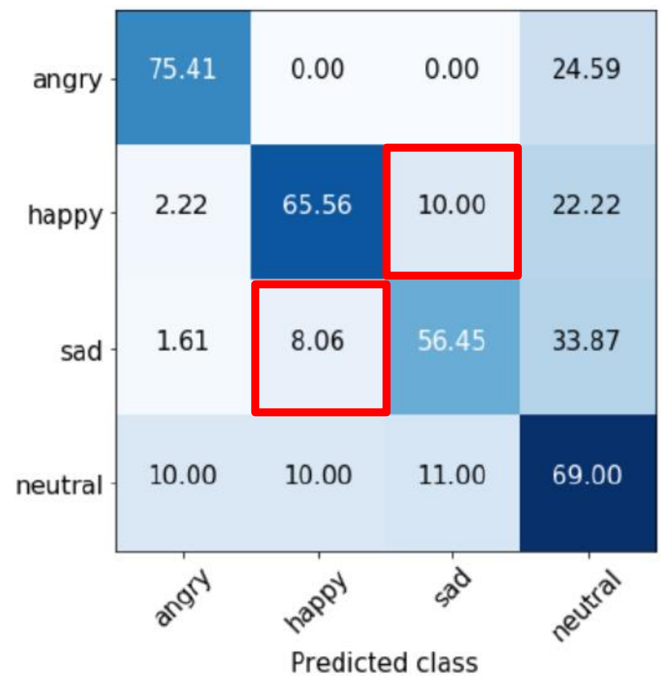
(b) text-BRE

Error Analysis

- **Text-BRE**
  - Incorrectly predicted instances of the "***happy***"as "**sad"** in 10%
  - even though these emotional states are opposites of one another



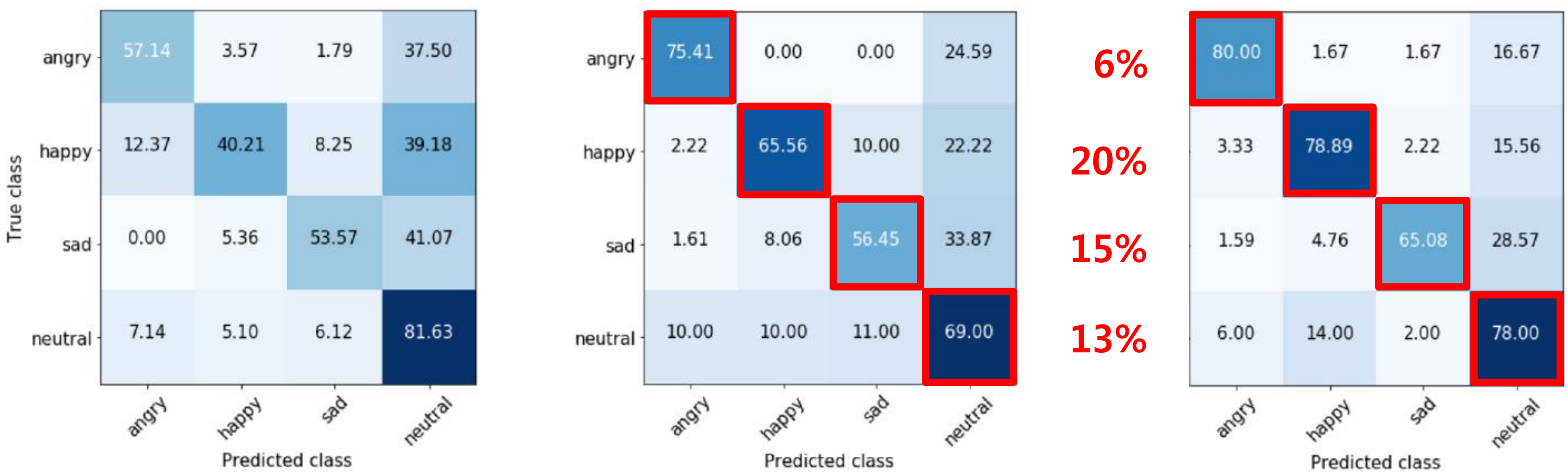
(a) audio-BRE



(b) text-BRE

Error Analysis

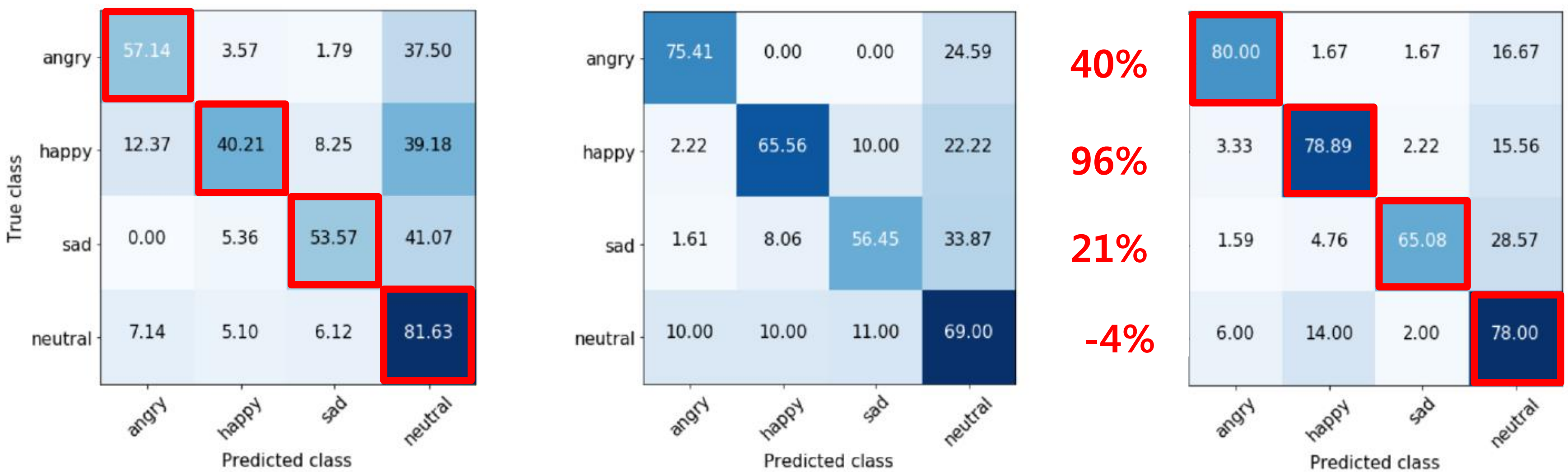
- **MHA-2**
  - Benefits from strengths of **audio-BRE** and **text-BRE**
  - Significant performance gain for all predictions (vs **text-BRE**)



(a) audio-BRE (b) text-BRE (c) MHA-2

Error Analysis

- **MHA-2**
  - Benefits from strengths of **audio-BRE** and **text-BRE**
  - Significant performance gain for all predictions (vs **audio-BRE**)



(a) audio-BRE (b) text-BRE (c) MHA-2

Conclusion

Consider NLP application?

→ Benefit From Large Data

Consider other application?

→ Benefit From NLP Technology

Thank you

mysmilesh@snu.ac.kr
http://david-yoon.github.io/