DialRC

Carnegie Mellon University
Language Technologies Institute

Structured Fusion Networks for Dialog

Shikib Mehri*, Tejas Srinivasan*, Maxine Eskenazi
Language Technologies Institute,
Carnegie Mellon University

Code: https://github.com/shikib/structured_fusion_networks

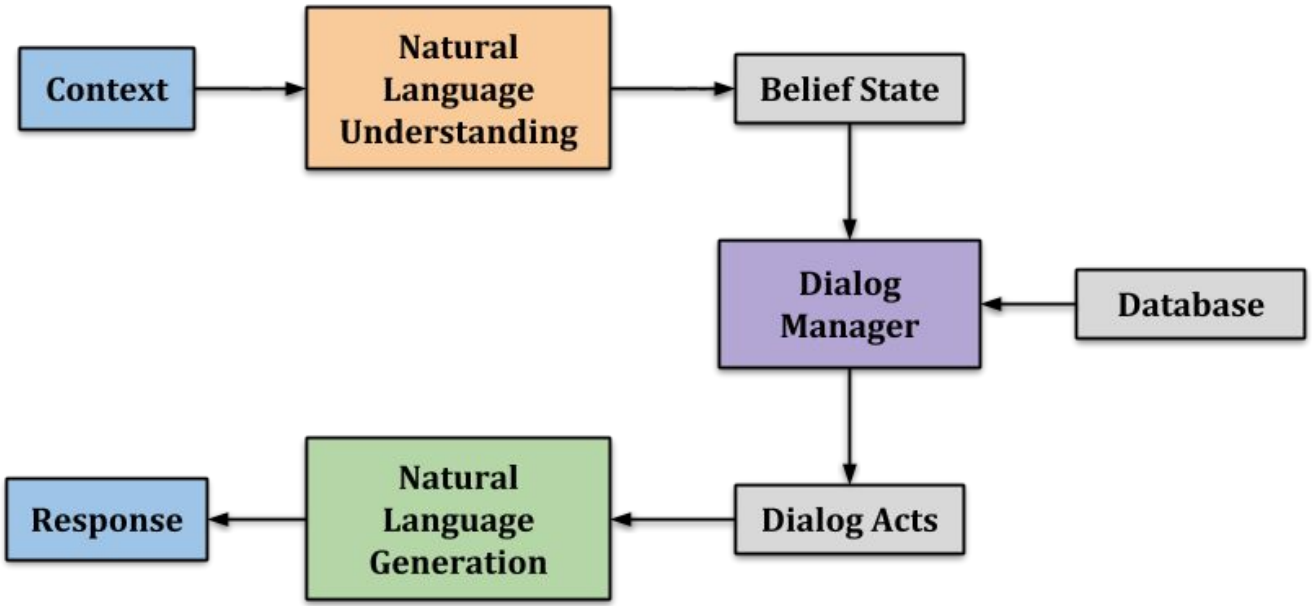
Motivation

Neural systems show **strong performance** but have shortcomings:

- **data-hungry** nature (Zhao and Eskenazi, 2018)
- inability to **generalize** (Mo et al., 2018)
- **lack of controllability** (Hu et al., 2017)
- **divergent behaviour** when tuned with RL (Lewis et al., 2017)

Traditional Pipeline Dialog Systems

Structured components facilitate effective **generalizability**, **interpretability** and **controllability**.



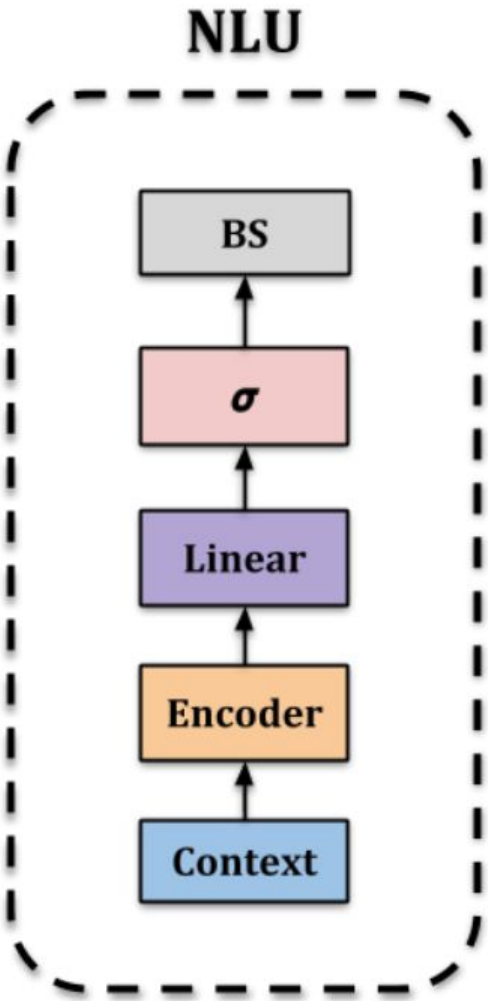
Feature	Traditional Dialog Systems	Neural Dialog Systems
Structured	✓	✗
Interpretable	✓	✗
Generalizable	✓	✗
Controllable	✓	✗
Higher-level reasoning/policy	✗	✓
Can learn from data	✗	✓

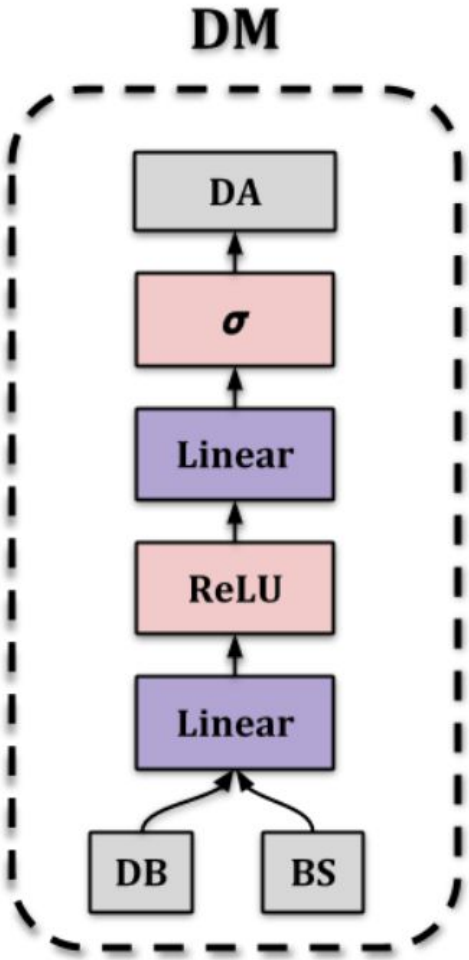
Why not combine the two approaches?

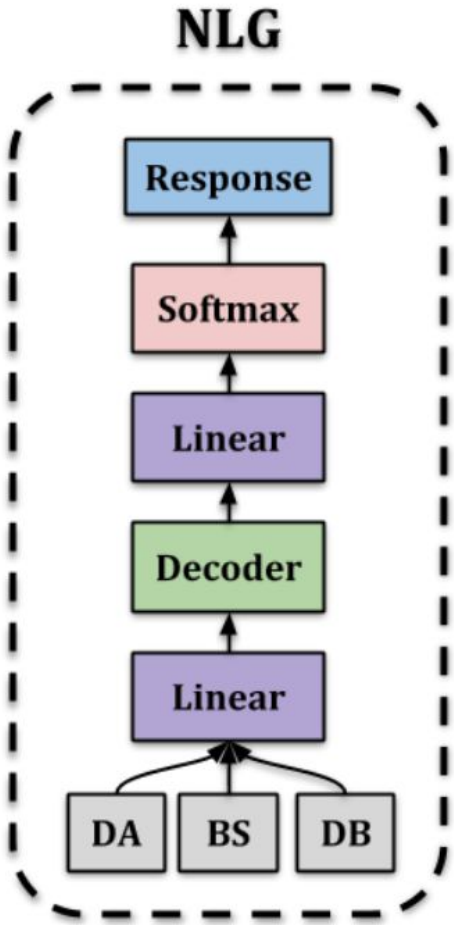
Neural Dialog Modules

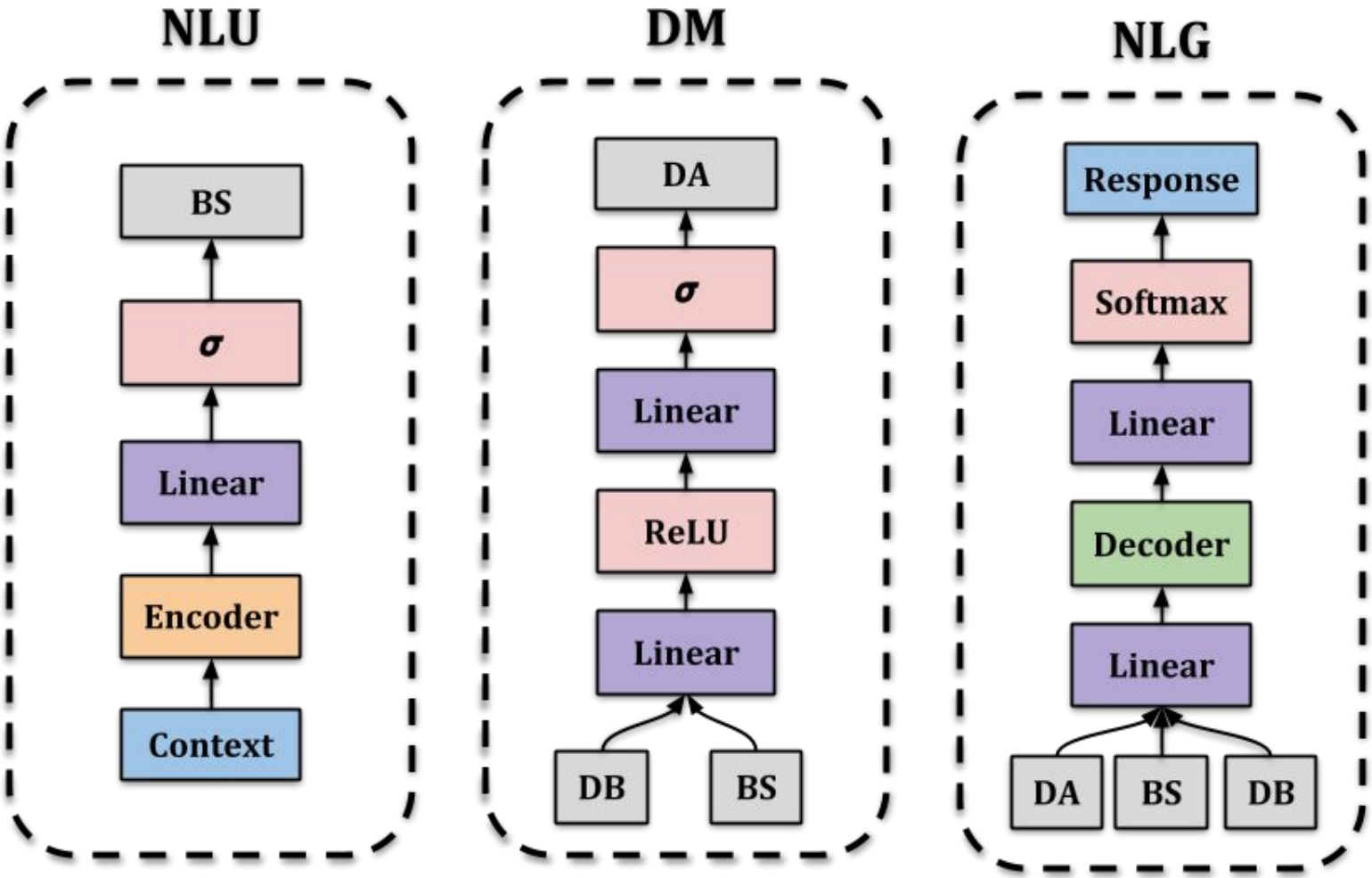
Using MultiWOZ (Budzianowski et al., 2018), define and train **neural dialog modules**

Natural Language Understanding (NLU)	dialog context → belief state
Dialog Manager (DM)	belief state → dialog acts for system response
Natural Language Generation (NLG)	dialog acts→ system response


NLU
BS
σ
Linear
Encoder
Context

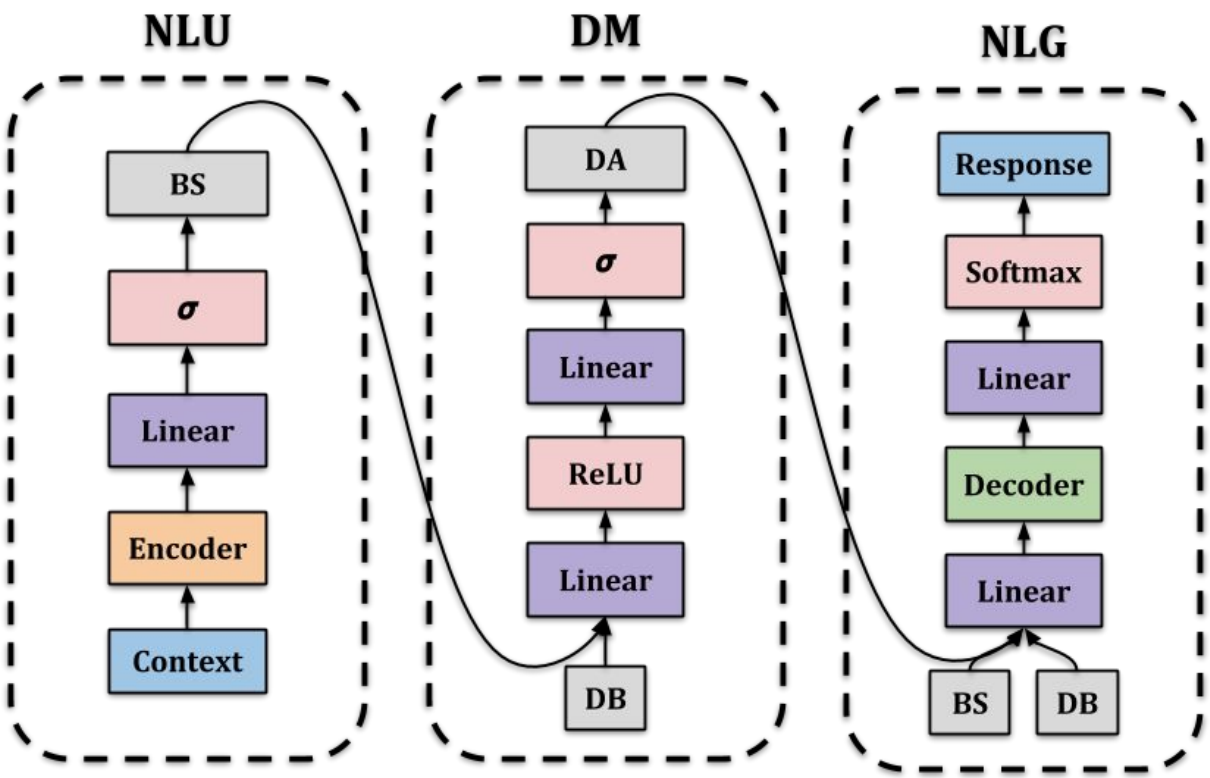

DM
DA
σ
Linear
ReLU
Linear
DB
BS


NLG
Response
Softmax
Linear
Decoder
Linear
DA
BS
DB


NLU
BS
σ
Linear
Encoder
Context
DM
DA
σ
Linear
ReLU
Linear
DB
BS
NLG
Response
Softmax
Linear
Decoder
Linear
DA
BS
DB

Naïve Fusion

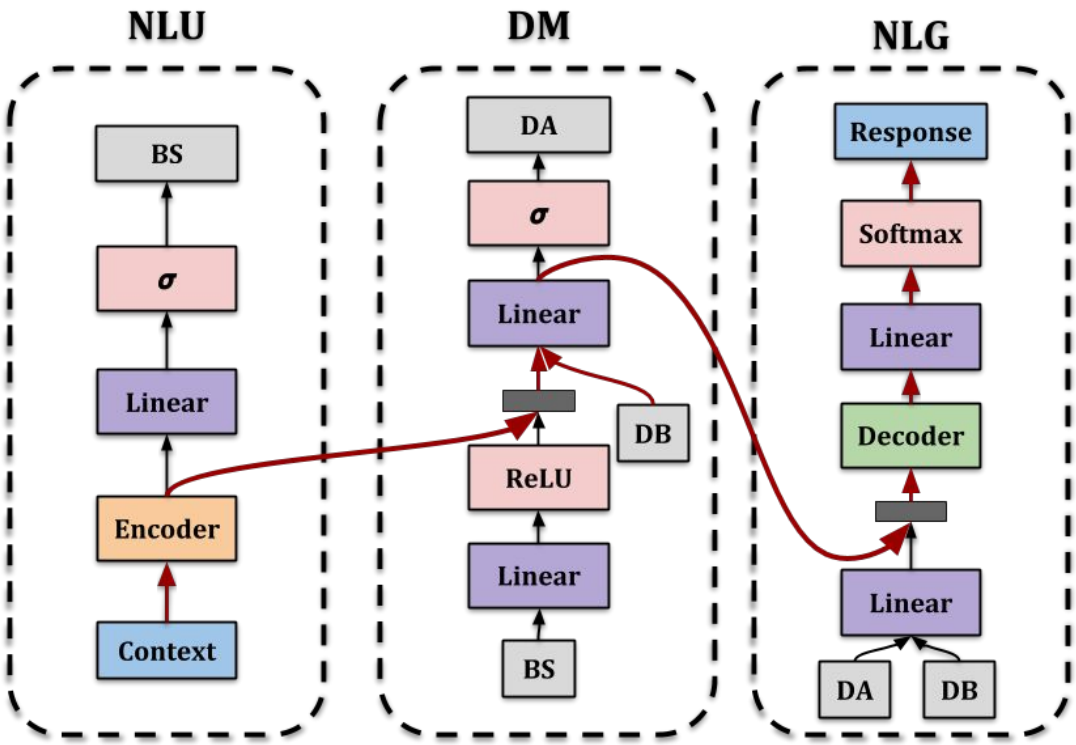
1. Train **neural dialog modules** independently
2. Combine them naively during inference
3. Give it a name → **Naïve Fusion**



Multi-Tasking

Simultaneously learn **dialog modules** and the final task of **dialog response generation**.

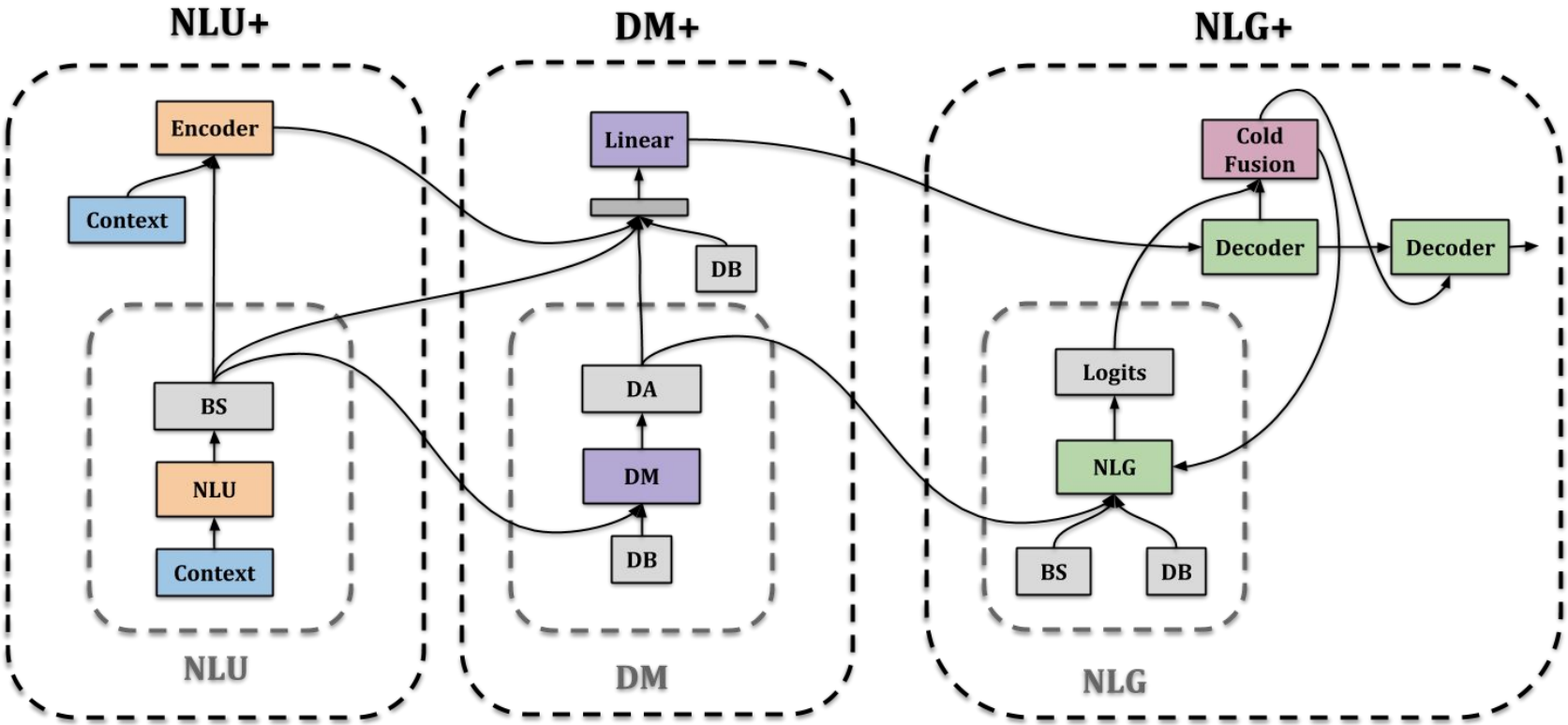
Sharing parameters results in more **structured components.**



Structured Fusion Networks

SFNs aim to learn a **higher-level model** on top of **pre-trained neural dialog modules**

Structured Fusion Networks

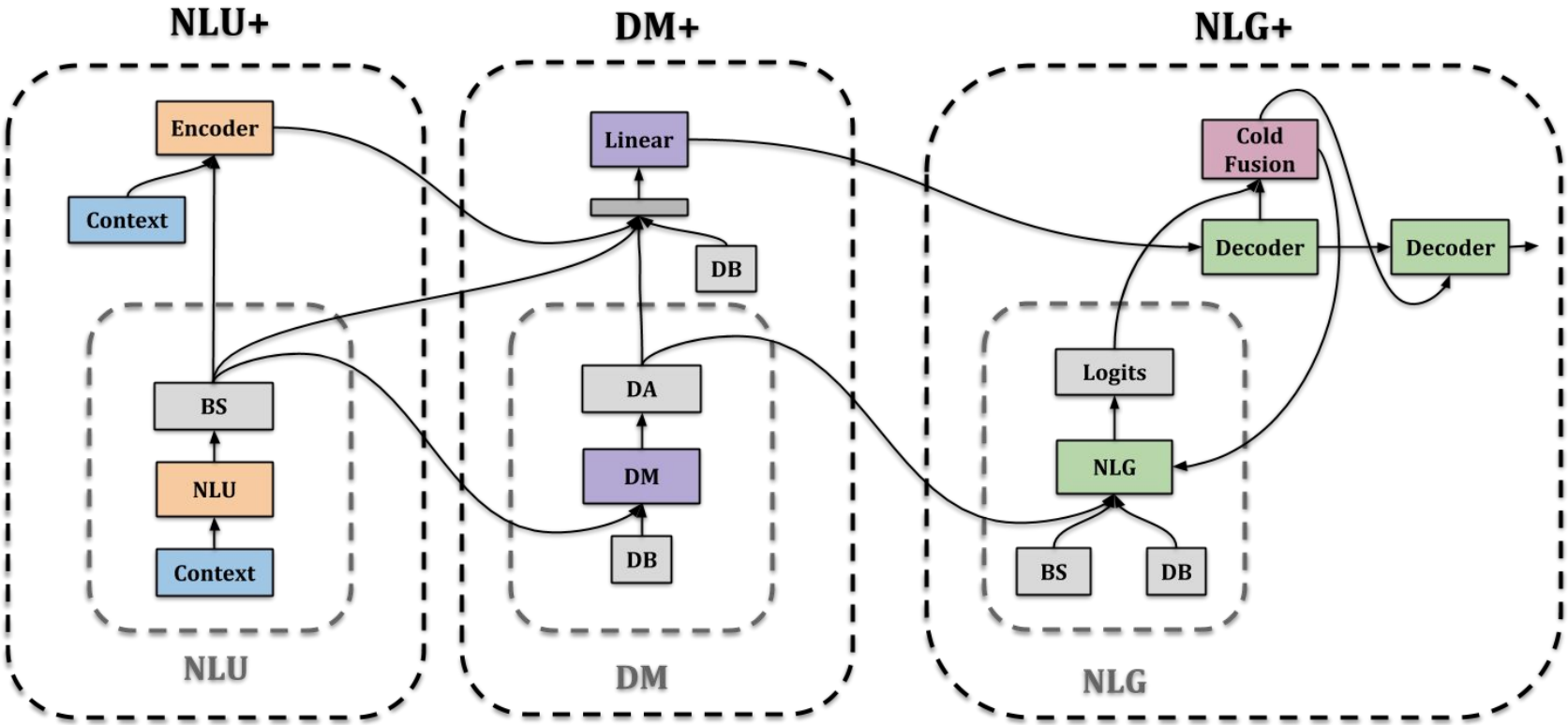


Structured Fusion Networks

SFNs aim to learn a **higher-level model** on top of **pre-trained neural dialog modules**

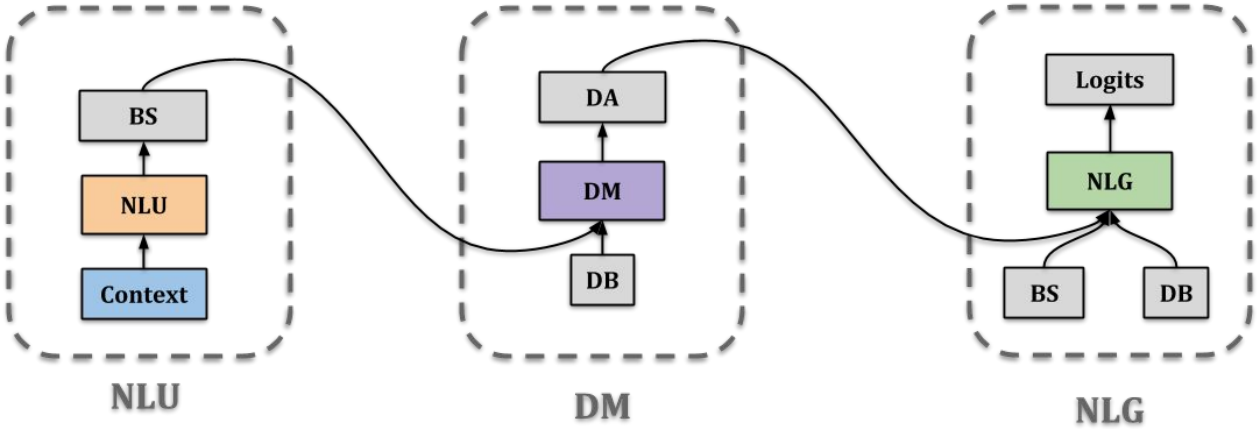
- Higher level model does not need to **re-learn** and **re-model** the dialog structure
- Instead can focus on **necessary abstract modelling**
  - **encoding** complex natural language
  - **policy modelling**
  - **generating language** conditioned on a latent representation

Structured Fusion Networks



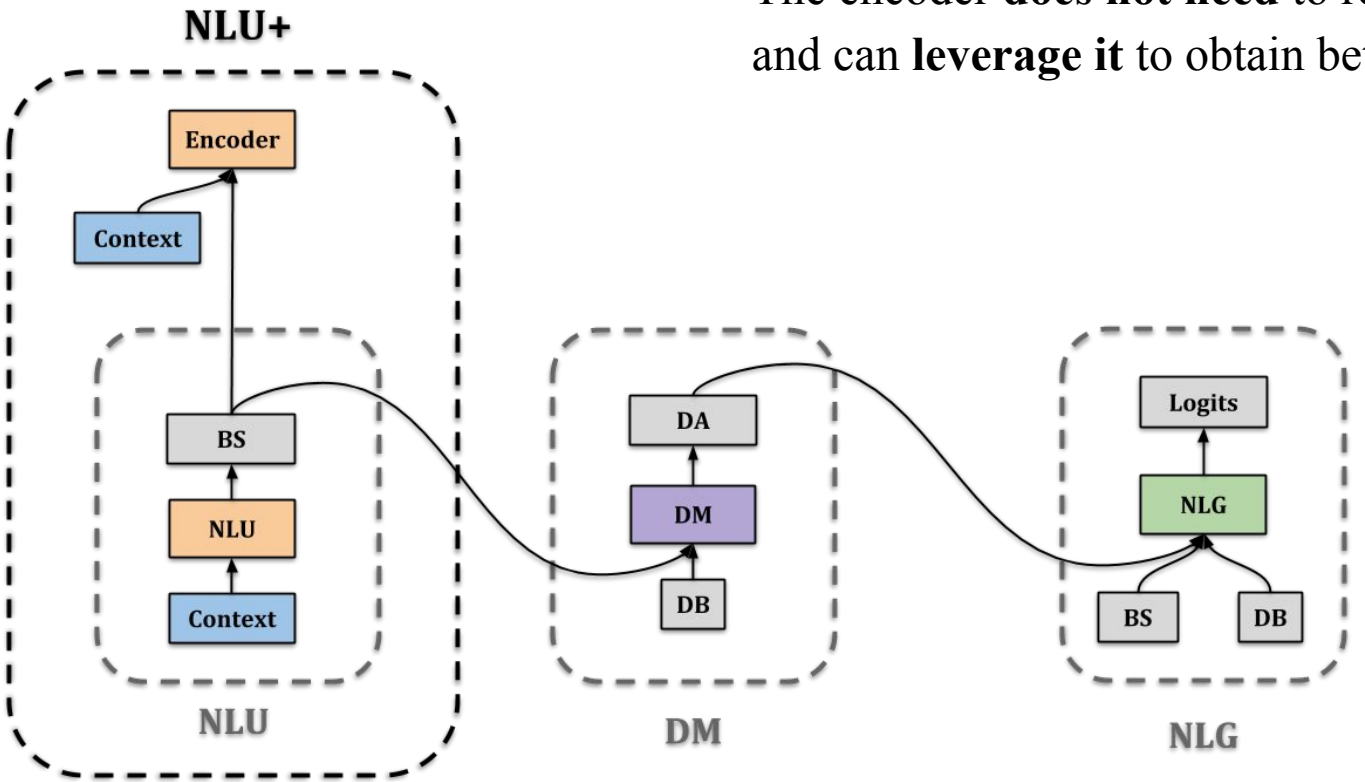
Dialog Modules

Start with **pre-trained** neural dialog modules

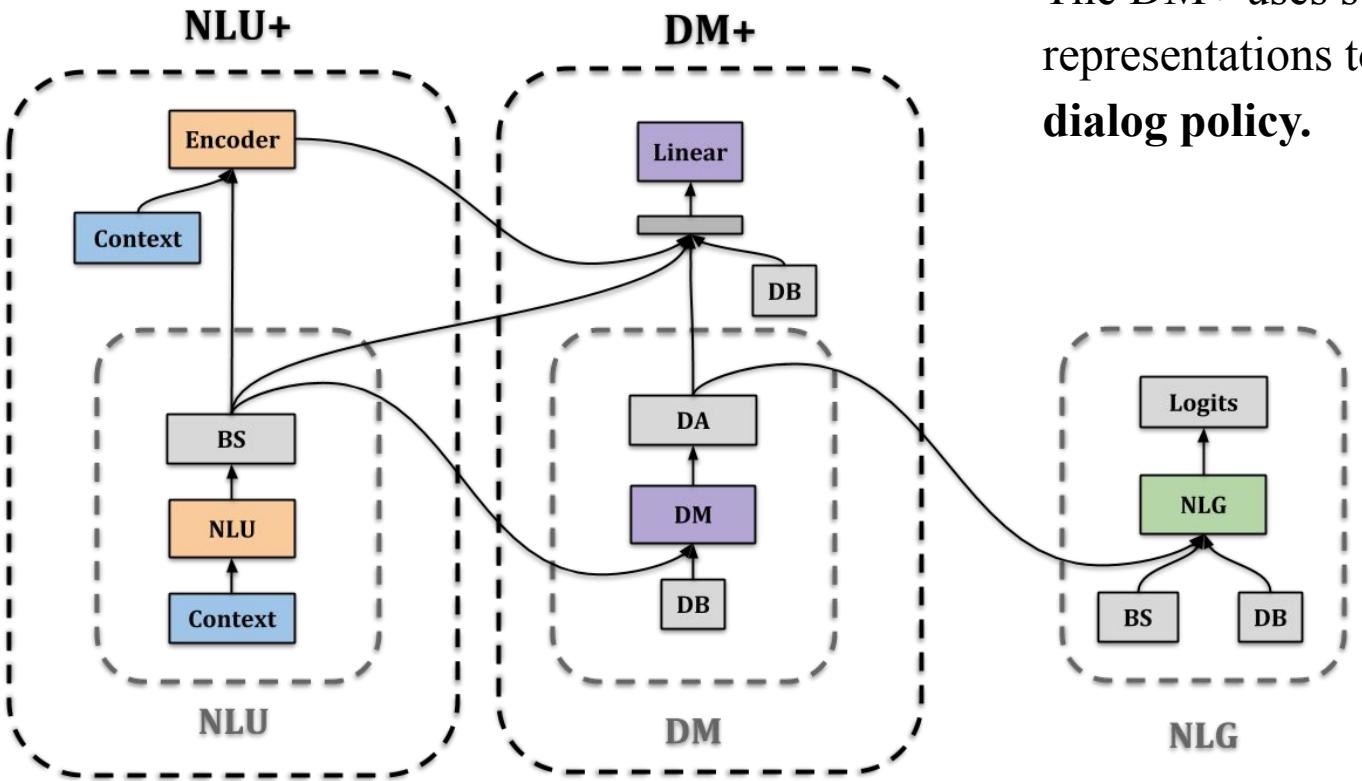


NLU+

The encoder **does not need** to re-learn the structure and can **leverage it** to obtain better encodings.

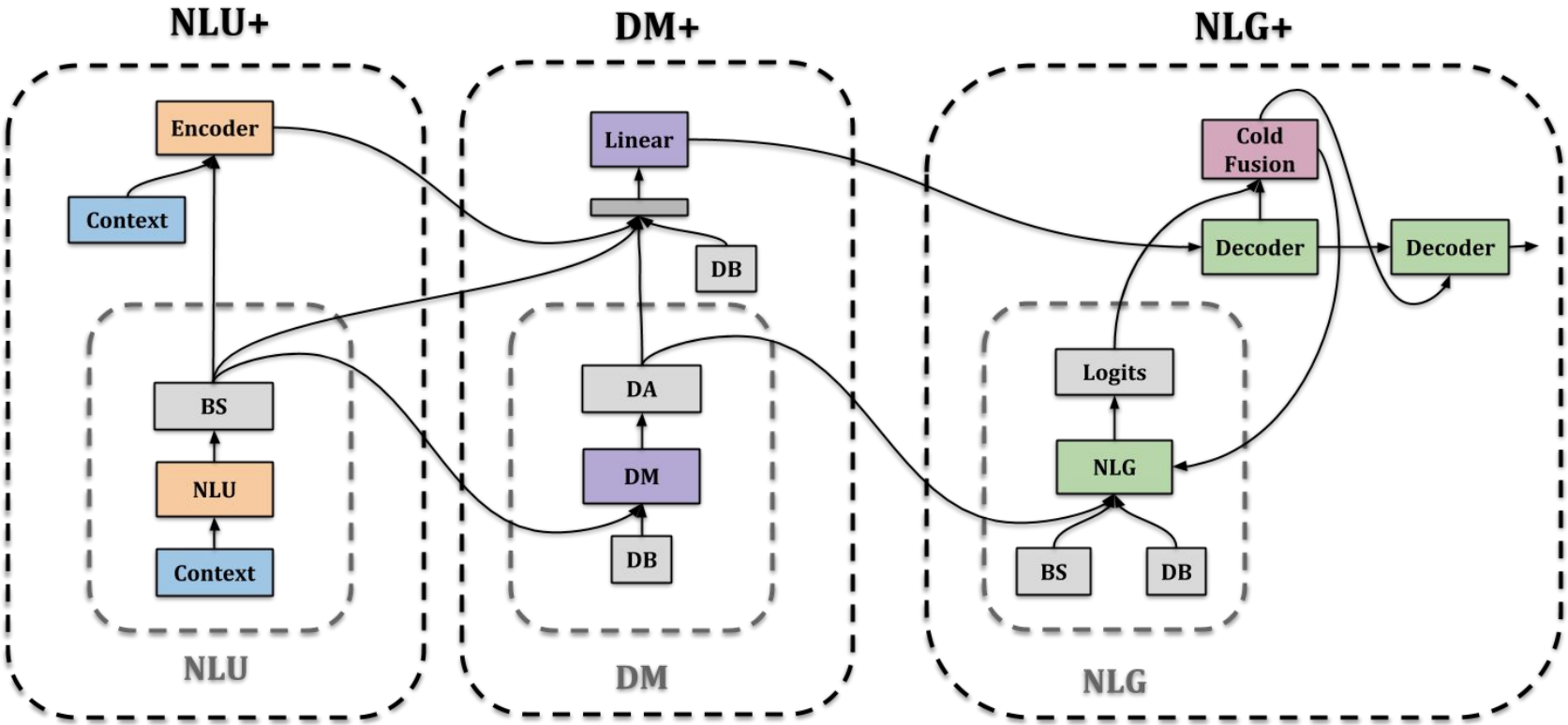


DM+



The DM+ uses structured representations to **explicitly model the dialog policy.**

NLG+


NLU+
DM+
NLG+
Encoder
Context
BS
NLU
Context
NLU
Linear
DB
DA
DM
DB
DM
Cold Fusion
Decoder
Decoder
Logits
NLG
BS
DB
NLG

NLG+

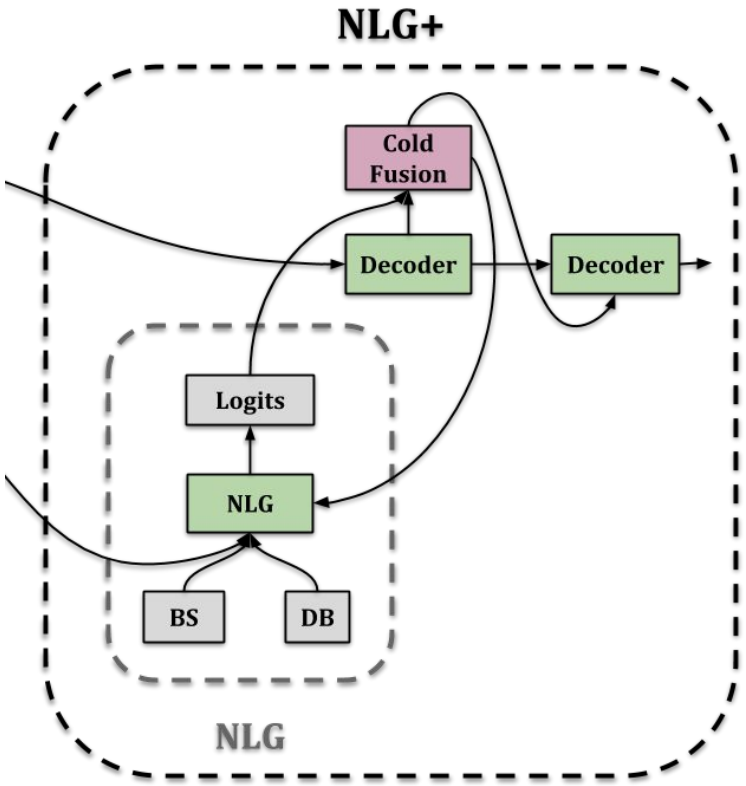
NLG+ relies on **Cold Fusion**.

NLG → sense of what the **next word** could be

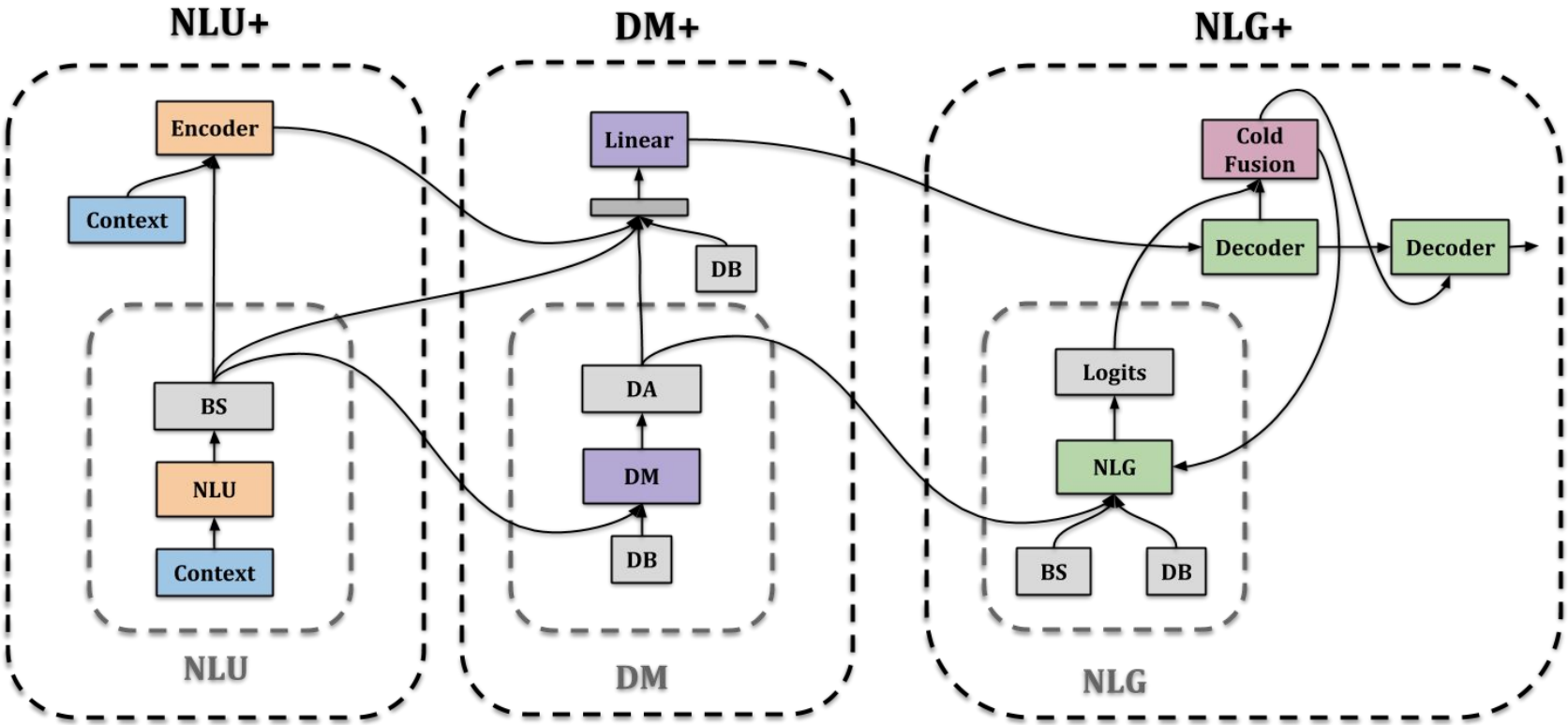
decoder → performs **higher-level reasoning**

ColdFusion →**combines** outputs

The outputs of the decoder are passed into the next time-step of the **NLG**.



Structured Fusion Networks



SFN Training

- Frozen modules
- Fine-tuned modules
- Multi-tasked modules

Experimental Setup

- MultiWOZ (Budzianowski et al., 2018)
  - Same hyperparameters
  - Use ground-truth belief state (oracle NLU)
- Evaluation
  - BLEU
  - Inform: *how often the system has provided the appropriate entities to the user*
  - Success: *how often the system answers all the requested attributes*
  - Combined = BLEU + 0.5*(Inform + Success)

Results

Model Name	BLEU	Inform	Success	Combined Score
Seq2Seq	**20.78**	61.40%	54.50%	78.73
Seq2Seq w/ Attn	20.36	66.50%	59.50%	83.36

Results

Model Name	BLEU	Inform	Success	Combined Score
Seq2Seq	**20.78**	61.40%	54.50%	78.73
Seq2Seq w/ Attn	20.36	66.50%	59.50%	83.36
Naive Fusion (Zero Shot)	7.55	70.30%	36.10%	60.75
Naive Fusion (Fine-Tuned)	16.39	74.70%	61.30%	84.39

Results

Model Name	BLEU	Inform	Success	Combined Score
Seq2Seq	**20.78**	61.40%	54.50%	78.73
Seq2Seq w/ Attn	20.36	66.50%	59.50%	83.36
Naive Fusion (Zero Shot)	7.55	70.30%	36.10%	60.75
Naive Fusion (Fine-Tuned)	16.39	74.70%	61.30%	84.39
Multi-Tasking	17.51	71.50%	57.30%	81.91

Results

Model Name	BLEU	Inform	Success	Combined Score
Seq2Seq	**20.78**	61.40%	54.50%	78.73
Seq2Seq w/ Attn	20.36	66.50%	59.50%	83.36
Naive Fusion (Zero Shot)	7.55	70.30%	36.10%	60.75
Naive Fusion (Fine-Tuned)	16.39	74.70%	61.30%	84.39
Multi-Tasking	17.51	71.50%	57.30%	81.91
SFN (Frozen)	17.53	65.80%	51.30%	76.08

Results

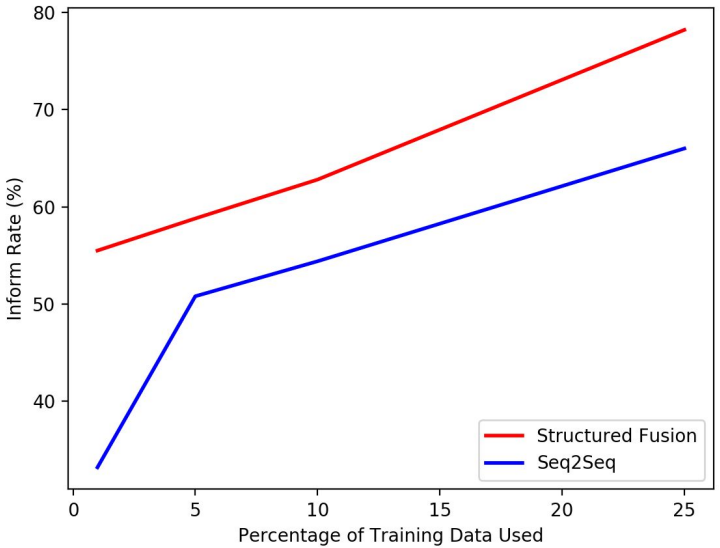
Model Name	BLEU	Inform	Success	Combined Score
Seq2Seq	**20.78**	61.40%	54.50%	78.73
Seq2Seq w/ Attn	20.36	66.50%	59.50%	83.36
Naive Fusion (Zero Shot)	7.55	70.30%	36.10%	60.75
Naive Fusion (Fine-Tuned)	16.39	74.70%	61.30%	84.39
Multi-Tasking	17.51	71.50%	57.30%	81.91
SFN (Frozen)	17.53	65.80%	51.30%	76.08
SFN (Fine-Tuned)	18.51	77.30%	**64.30%**	**89.31**

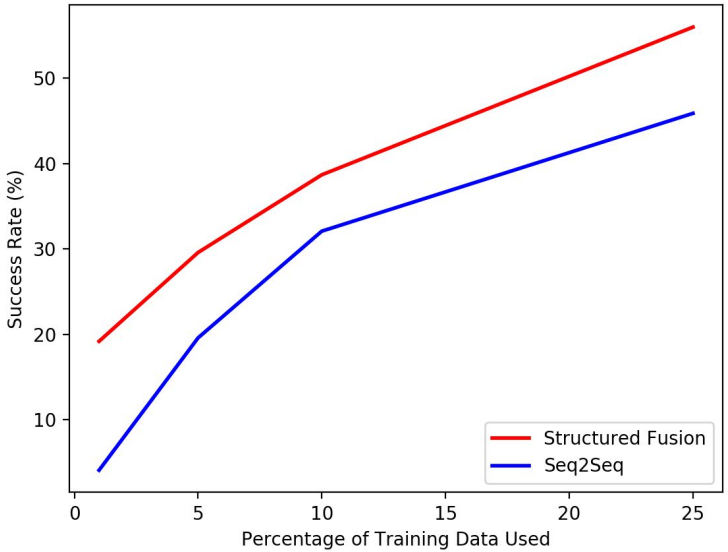
Results

Model Name	BLEU	Inform	Success	Combined Score
Seq2Seq	**20.78**	61.40%	54.50%	78.73
Seq2Seq w/ Attn	20.36	66.50%	59.50%	83.36
Naive Fusion (Zero Shot)	7.55	70.30%	36.10%	60.75
Naive Fusion (Fine-Tuned)	16.39	74.70%	61.30%	84.39
Multi-Tasking	17.51	71.50%	57.30%	81.91
SFN (Frozen)	17.53	65.80%	51.30%	76.08
SFN (Fine-Tuned)	18.51	77.30%	**64.30%**	**89.31**
SFN (Multi-tasked)	16.70	**80.40%**	63.60%	88.71

Limited Data

The added structure should result in **less data-hungry models**. We compare **Seq2Seq** and **SFN** when using 1%, 5%, 10% and 25% of the training data.





Domain Generalizability

The added structure should result in **more generalizable models**. We compare **Seq2Seq** and **SFN** on their in-domain (restaurant) performance, using **2000 out-of-domain** examples and **50 in-domain** examples**.**

Model Name	BLEU	Inform	Success	Combined Score
Seq2Seq	**10.22**	35.65%	1.30%	28.70
SFN	7.44	**47.17%**	**2.17%**	**32.11**

Divergent Behaviour with RL

Training **generative dialog models** with **RL** often results in divergent behavior and degenerate output (Lewis et al., 2017, Zhou et al., 2019)

Implicit Language Model

Standard decoders have the issue of the **implicit language model**. The decoder simultaneously learns to follow some policy and model language.

In image captioning (Wang et al., 2016), the implicit language model **overwhelms** the decoder.

Fine-tuning dialog models with RL causes it to **unlearn** the implicit language model.

But SFN's have an **explicit LM**

SFN + Reinforcement Learning

We pre-train an SFN with supervised learning, we then **freeze the dialog modules** and **fine-tune only the higher-level model** with a reward of ***Inform+Success***

This way, we use RL to optimize the higher-level model for some **dialog strategy** while also maintaining the **structured nature** of the dialog modules

Model Name	BLEU	Inform	Success	Combined Score
Seq2Seq + RL (Zhao et al. 2019)	1.40	80.50%	**79.07%**	81.19
LiteAttnCat + RL (Zhao et al. 2019)	12.80	**82.78%**	**79.20%**	**93.79**
SFN (Frozen Modules) + **RL**	**16.34**	**82.70%**	72.10%	**93.74**

Results

Model Name	BLEU	Inform	Success	Combined Score
SFN (Fine-Tuned)	**18.51**	77.30%	64.30%	89.31
SFN (Multi-tasked)	16.70	80.40%	63.60%	88.71
Seq2Seq + RL (Zhao et al. 2019)	1.40	80.50%	**79.07%**	81.19
LiteAttnCat + RL (Zhao et al. 2019)	12.80	**82.78%**	**79.20%**	**93.79**
SFN (Frozen Modules) + **RL**	16.34	**82.70%**	72.10%	**93.74**

Results

Model Name	BLEU	Inform	Success	Combined Score
SFN (Fine-Tuned)	18.51	77.30%	64.30%	89.31
SFN (Multi-tasked)	16.70	80.40%	63.60%	88.71
Seq2Seq + RL (Zhao et al. 2019)	1.40	80.50%	**79.07%**	81.19
LiteAttnCat + RL (Zhao et al. 2019)	12.80	**82.78%**	**79.20%**	93.79
SFN (Frozen Modules) + **RL**	16.34	**82.70%**	72.10%	93.74
HDSA (Chen et al., 2019)*	**23.60**	**82.90%**	68.90%	**99.50**

* Released after our paper was in-review. Room for combination.

Human Evaluation

Asked **AMT workers** to read the dialog context and rate several responses on a scale of 1-5 on **appropriateness**.

Model Name	Average Rating	≥ 4	≥ 5
Seq2Seq	3.00	40.21%	9.61%
SFN	3.02	**44.84%**	11.03%
SFN + RL	**3.12**	**44.84%**	**16.01%**
Human Ground Truth	3.76	59.75%	34.88%

DialRC

Carnegie Mellon University
Language Technologies Institute

Multi-Granularity Representations of Dialog

Shikib Mehri, Maxine Eskenazi
Language Technologies Institute,
Carnegie Mellon University

Code: https://github.com/shikib/structured_fusion_networks

Motivation

Recent research has tried to produce **general latent representations of language** (ELMo, BERT, GPT-2 … etc.)

Why is it so hard to get these representations to work well for dialog?

1. **Domain difference**
2. **LM objectives do not necessarily capture properties of dialog**

Goal: strong and general representations of dialog

Motivation

Goal: strong and general representations of dialog

- Large pre-trained models: **general** but **not strong** (at dialog)
- Task-specific models: **strong** but **not general** (won't generalize to other tasks)

Generality?

Text → Latent Representation *results in a **loss of information***

- Neural models will always **look for a shortcut**
  - If they can fall into a local optima by simple pattern matching, they will
  - **Well-formulated tasks** result in good representations
- Impossible to construct a *one size fits all* representation using a single task
  - Representation will focus on the **average example**

Generality

Example: imagine we are using a sentence similarity as a pre-training task. Let's think about the types of representations we would get.

Case 1: Train on very similar sentences

- *The cat in the hat ran into the room*
- *The cat in the hat strolled into the room*

We would get very **granular** representations. Maybe the model will learn to look at **keywords** and construct strong representations of **actions**.

Generality

Example: imagine we are using a sentence similarity as a pre-training task. Let's think about the types of representations we would get.

Case 2: Train on very different sentences

- ➢ *The cat in the hat ran into the room*
- ➢ *He was the first man to walk on the moon*

We would get very **broad** representations. Maybe the model will learn to look at **topic** and construct strong representations of **domain/topic**.

Proposed solution

Problem

Neural models **look for shortcuts** and **fit to the average of the training data**.

Different **granularities of representation** are difficult to capture.

Proposed solution

Formulate a mechanism of **learning multiple granularities of representation**, then combine the different representations into a **multi-granularity representation.**

Dialog Retrieval

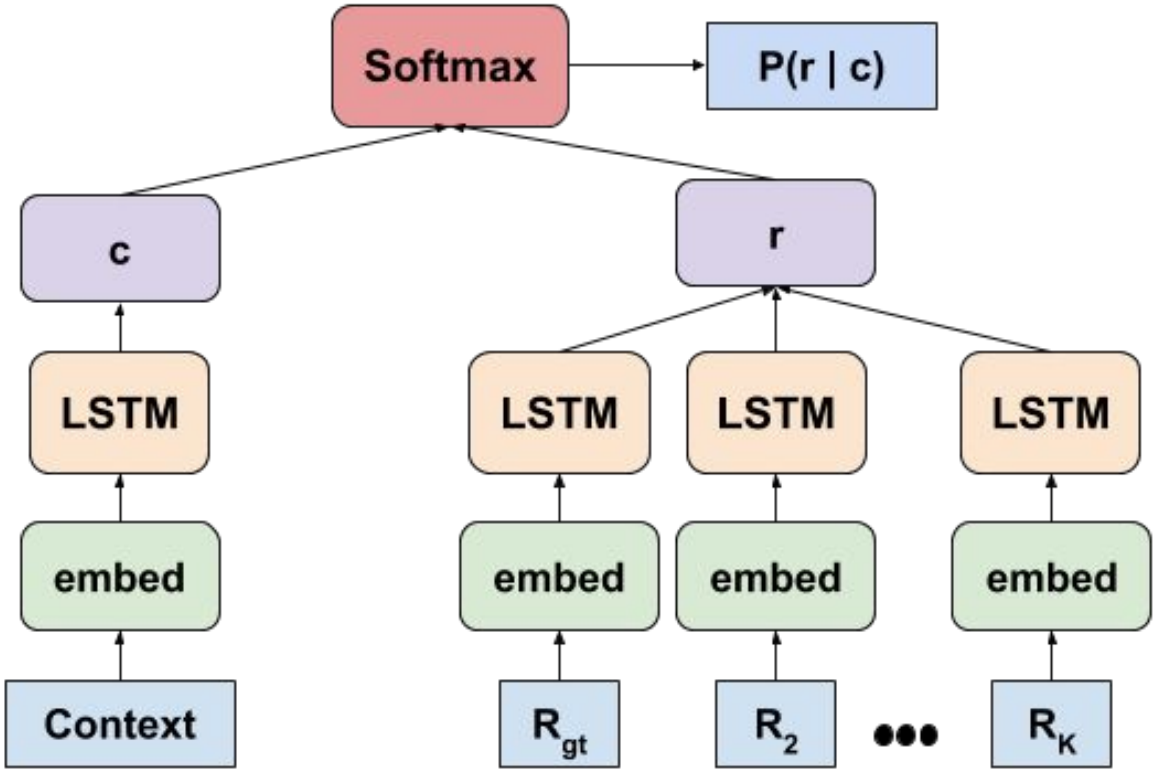
Input:

- dialog context (history) consisting of utterances
- set of candidate responses (with one correct response)

Task: Retrieve the correct response, using the dialog context, from the set of candidate responses.

Data: MultiWoz (Budzianowski et al., 2018) & Ubuntu Dialog Corpus (Lowe et al., 2015)

Baseline Model



Multi-Granularity

Negative candidates influence **granularity** of **representations**

similar candidates → granular representations

distant candidates → abstract representations

Multi-Granularity

Negative candidates influence **granularity** of **representations**

1. Construct a **similarity measure**
2. Construct **candidate sets** of different distances
3. Train M models on different distances of candidate sets. Each model will capture a **different granularity** of representation.

Similarity Measure

1. Train a retrieval model
2. Produce latent representations of each response
3. Cosine similarity

Dialog Context:

USER: i need an expensive place to dine in the center of town .

SYS: what type of food are you wanting at this restaurant ?

USER: type of food does not matter to me .

Ground-Truth Response:

SYS: yes there are 33 such restaurant -s . would you like me to book 1 for you ?

Closest Neg. Cand. (for Highest Granularity Model):

SYS: i have 5 indian restaurant -s in the west . if i may , might i suggest the rice boat . i can reserve a table for you if you would like .

2nd Closest Neg. Cand. (for 2nd Highest Granularity Model):

SYS: how about the avalon ?

3rd Closest Neg. Cand. (for 3nd Highest Granularity Model):

SYS: i would be happy to book that for you .are you travelling alone ?

2nd Furthest Neg. Cand. (for 2nd Lowest Granularity Model):

SYS: i apologize , there s been a misunderstanding . before we can book your taxi , we need to know which college . there are 18 colleges . what area are you looking for ?

Furthest Granularity Neg. Cand (for Lowest Granularity Model):

SYS: i need to know when you want to leave and arrival time in order for me to order you a taxi .

Multi-Granularity Training

Train 5 retrieval models on each of the candidate sets.

Closer candidate sets → Granular representations

Farther candidate sets → Abstract representations

Ensemble models after training

Retrieval Metrics

- R_k@1 *Accuracy of selecting the ground-truth response from k negative candidates*
- MRR *Mean Reciprocal Rank*

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank_i}$$

Retrieval Results (MultiWOZ)

Model Name	MRR	R_{20}@1
Dual Encoder	79.55	66.13%
Dual Encoder Ensemble (5)	81.53	69.47**%**
Multi-Granularity (5)	**82.74**	**72.18%**

Retrieval Results (Ubuntu)

Model Name	MRR	$R_{10}@1$	$R_2@1$
Dual Encoder (Lowe et al., 2015)	-	63.8%	90.1%
DL2R (Yan et al., 2016)	-	62.6%	89.9%
SMN (Wu et al., 2016)	-	72.6%	92.6%
DAM (Zhou et al., 2018)	**-**	**76.7%**	**93.8%**
Dual Encoder	76.84	63.6%	90.9%
Dual Encoder Ensemble (5)	78.91	66.9%	91.7%
Multi-Granularity (5)	**80.10**	**68.7%**	**91.9%**

Retrieval Results (Ubuntu) + DAM

Model Name	MRR	$R_{10}@1$	$R_2@1$
Dual Encoder	76.84	63.6%	90.9%
Dual Encoder Ensemble (5)	78.91	66.9%	91.7%
Multi-Granularity (5)	80.10	68.7%	91.9%
DAM (re-trained)	83.74	74.5%	93.1%
DAM Ensemble (5)	84.03	75.0%	93.3%
DAM Multi-Granularity (5)	**84.26**	**75.3%**	**93.5%**

Are we really learning different granularities?

Performance on retrieval shows we learn **more diverse** models, but are we really learning different granularities of representation?

- Freeze the model
- Use pre-trained representations to train on downstream tasks of different granularities
  - Bag of Words prediction (*high granularity task)*
  - Next Dialog Act prediction (*high abstraction task*)

Granularity Analysis

Model Name	BoW (F-1)	DA (F-1)
Highest Abstraction	57.00	**19.24**
2nd Highest Abstraction	57.69	19.14
Medium	58.49	18.31
2nd Highest Granularity	58.38	16.88
Highest Granularity	**59.43**	15.46

Generalizable Representation (No Fine-tuning)

Model Name	BoW (F-1)	DA (F-1)
Dual Encoder	60.13	19.09
Dual Encoder Ensemble (5)	64.11	22.39
Multi Granularity (5)	**67.51**	**22.85**
Random Init + Fine-Tuned	90.33	28.75

Generalizable Representation (Fine-tuning)

Model Name	DA (F-1)
Random Init	28.75
Dual Encoder	32.63
Dual Encoder Ensemble (5)	31.71
Multi Granularity (5)	**33.46**

Takeaways

Want **strong** and **general** representations of dialog

Strong: *Train on dialog data for a dialog task*

General: *Learn multiple granularities of representation, to avoid fitting to the mean of the data.*

Future Work (MGT)

- Apply multi-granularity training to other tasks
- More sophisticated similarity measure/model combination
- Generalize to language generation
- Learn representations along several different axes (domain, styles, intents)
  - Without explicit specification

Future Work (SFN)

- Generalize to open-domain
- Explore controllability with structured components
- Analyze impacts of different components on model quality
- Combine with recent advances on MultiWOZ dataset

Code available at (or scan the QR code)
https://github.com/shikib/structured_fusion_networks

Thank you for your attention.

Model	BLEU	Inform	Success	Comb.
GT	**18.51**	**77.30%**	**64.30%**	**89.31**
Pred	16.88	73.80%	58.60%	83.04
Sum	15.93	72.90%	60.80%	82.78
Linear	15.42	66.80%	54.80%	76.22

Table 4: Results of the domain transfer experiment comparing sequence-to-sequence and Structured Fusion Networks. All bold-face results are statistically significant ($p < 0.01$).

(1) SFN fine-tuned with RL **consistently provides more attribute information.** It provides at least one attribute in every example response, for a total of 14 total attributes across the four examples. This, along with the high Success score of this model, is a consequence of the reinforcement learning fine-tuning which directly optimizes the Success score and rewards the model for providing the correct attributes.

(2) Seq2Seq **produces more generic responses**. In the second and fourth examples, the outputs produced by Seq2Seq are generic and are unrelated to the context. This suggests that the Seq2Seq model has an overwhelming implicit language model, which produces generic and dull responses. On the other hand, the explicit structure in SFN mitigates the effect of the implicit language model by relying on an *explicit* language model (in the form of the NLG).

(3) Seq2Seq **requests attributes which the user has already provided.** In the first and third example, the Seq2Seq output is requesting parameters that the user explicitly provided. This highlights the lack of structured representations in the Seq2Seq model. SFN which explicitly models the structure of the input effectively understands and captures the information provided in the context.

Example 1
Dialog Context: USER: hello . i am looking for a [value_pricerange] place to eat in the [value_area] . can you help me ?
Sequence-to-Sequence Response: i have [value_count] options for you . do you have a preference on price range ?
Structured Fusion Network Response: [restaurant_name] is a [value_pricerange] -ly priced [value_food] restaurant in the [value_area] . would you like to make a reservation ?
Structured Fusion Network w/ Reinforcement Learning Response: [restaurant_name] is a [value_pricerange] [value_food] restaurant in the [value_area] . it is located at [restaurant_address] . the postcode is [restaurant_postcode] . the phone number is [restaurant_phone]
Example 2
Dialog Context: . . . USER: i want to book this hotel for the same group of people and [value_count] nights starting from [value_day] . i'll need a reference number please . SYSTEM: your reservation is confirmed and the reference number is ## USER: i am sorry ? were you able to book the express by holiday inn for [value_count] people ?
Sequence-to-Sequence Response: you are welcome . have a great day !
Structured Fusion Network Response: you are all set . your reference number is [hotel_reference] . can i help you with anything else ?
Structured Fusion Network w/ Reinforcement Learning Response: your booking was successful . the reference number is [hotel_reference] .

Cold Fusion

$$h_t^{NLG} = DNN(l_t^{NLG}) \quad (7)$$

$$g_t = \sigma(W[s_t; h_t^{NLG}] + b) \quad (8)$$

$$s_t^{CF} = [s_t; g_t \circ h_t^{NLG}] \quad (9)$$

$$y_t = softmax(DNN(s_t^{CF})) \quad (10)$$