Statistical Dialogue Systems

NPFL099 Statistické Dialogové systémy

9. End-to-end systems (2)

Ondřej Dušek & Vojtěch Hudeček

http://ufal.cz/npfl099

12. 12. 2019

(Shu et al., 2019) https://www.aclweb.org/anthology/W19-5922/

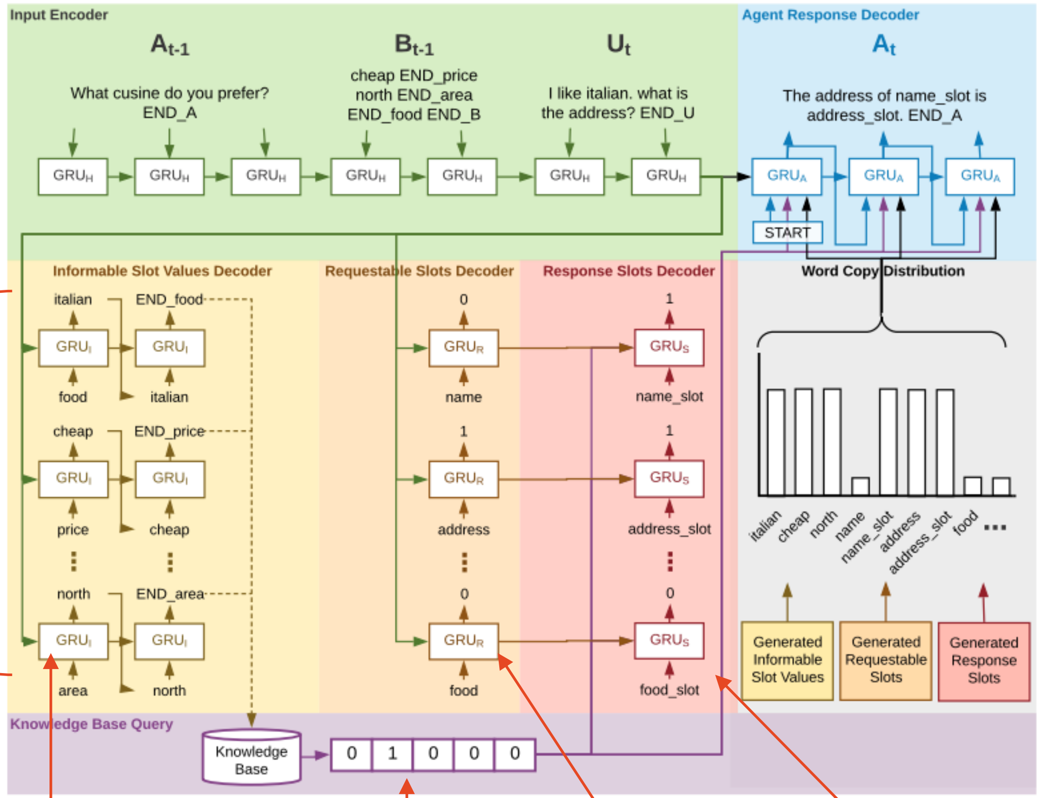
Sequicity + explicit state

- the same context encoder as Sequicity
- state decoder:
  - individual slots decoded separately
    - **prevents decoding invalid states**
  - the same decoder run for each slot
  - informable:
    - decode values, seq2seq way
  - requestable:
    - classify 0/1 if user requested
- response generation:
  - 1st step – classify which slots to include
  - then seq2seq delexicalized generation

prev. system output

prev. state

user input



tied weights: only the start symbol differs

1-hot 0/1/2/3/4/5 or more results

seq2seq decoder for user-specified slot values

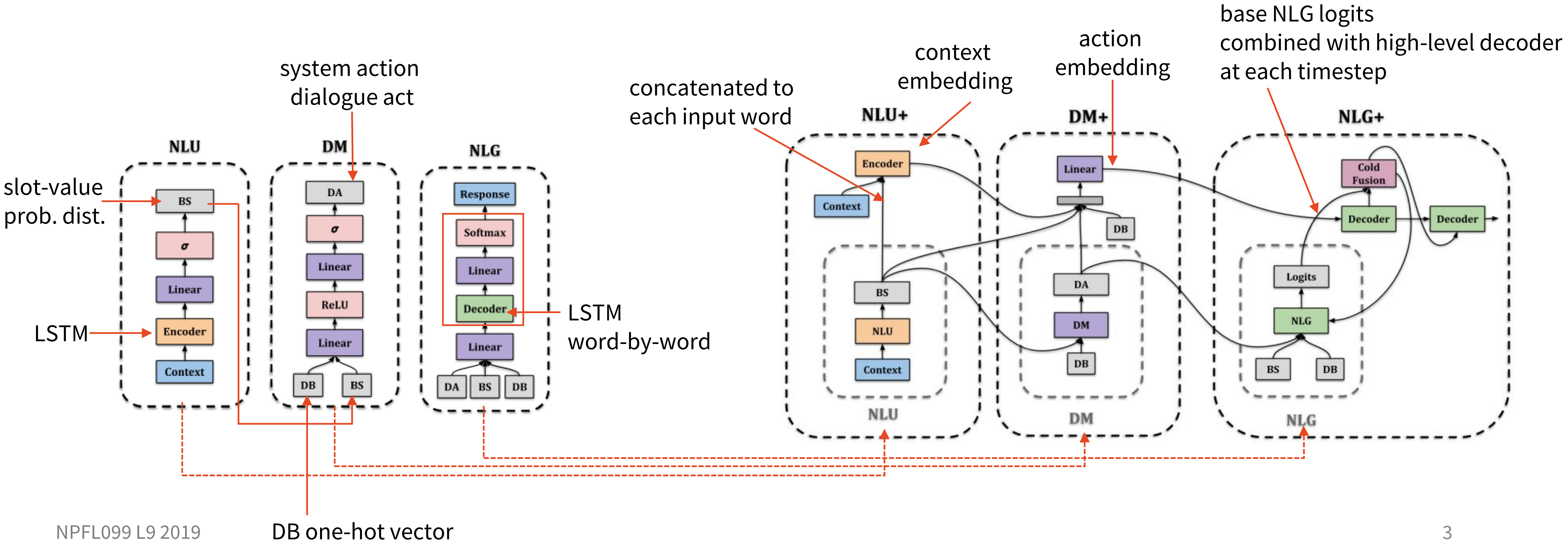
binary classifier: should slot X be in response?

binary classifier: did user request slot X?

Structured Fusion Nets:
End-to-end on top of individual modules

(Mehri et al., 2019)
https://www.aclweb.org/anthology/W19-5921/

- 1st step: optimize separate NLU/DM/NLG modules
- 2nd step: optimize end-to-end network over the outputs of modules



Structured Fusion Nets

(Mehri et al., 2019)
https://www.aclweb.org/anthology/W19-5921/

- high-level module on top of NLU/DM/NLG modules works better than just joining, even with joint optimization
- modules can be fine-tuned (end-to-end differentiable)
  - this helps in either case (with modules only or high-level network)
  - multi-task learning doesn't help more (alternating fine-tuning with module-specific tasks)
- RL: only high-level
  - this way the base generator maintains fluency
  - BLEU OK & success much higher

% dialogues where appropriate entity was provided

Model	BLEU	Inform	Success
Supervised Learning			
Seq2Seq (Budzianowski et al., 2018)	18.80	71.29%	60.29%
Seq2Seq w/ Attention (Budzianowski et al., 2018)	18.90	71.33%	60.96%
Seq2Seq (Ours)	20.78	61.40%	54.50%
Seq2Seq w/ Attention (ours)	20.36	66.50%	59.50%
Naïve Fusion (Zero-Shot)	7.55	70.30%	36.10%
Naïve Fusion (Fine-tuned Modules)	16.39	66.50%	59.50%
Multitasking	17.51	71.50%	57.30%
Structured Fusion (Frozen Modules)	17.53	65.80%	51.30%
Structured Fusion (Fine-tuned Modules)	18.51	77.30%	64.30%
Structured Fusion (Multitasked Modules)	16.70	80.40%	63.60%
Reinforcement Learning			
Structured Fusion (Frozen Modules) + RL	**16.34**	**82.70%**	72.10%

modules only

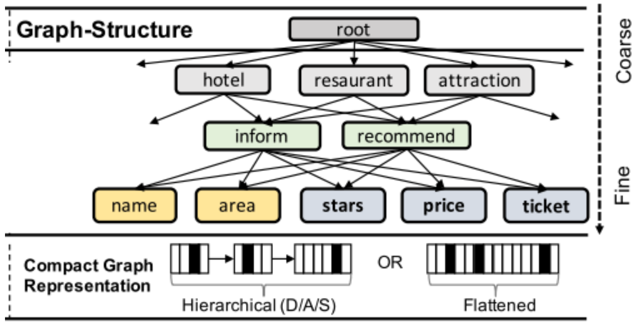
with high-level structure

% dialogues where system also provided all requested slots

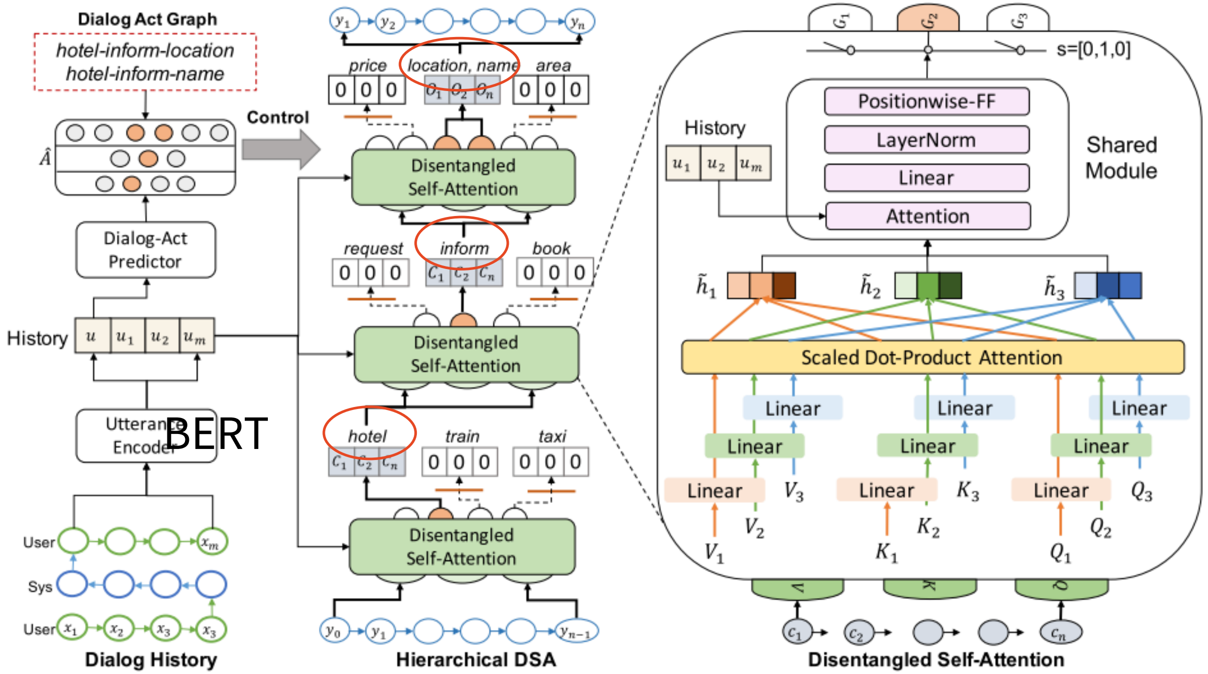
MultiWOZ (multi-domain data)

DA-based self-attention

- DAs represented as a graph
  - 3-level: domains – intents – slots



- ignores DB & tracker
  - uses ground truth from data
- NLU:
  - BERT over all history tokens
  - feed-forward/attention + sigmoid
  - predict domains-intents-slots graph
- Decoder: modified self-attention
  - optimized separately
  - gated sum instead of concatenation
    - gating follows predicted DA graph
  - delexicalized – DB & tracker provide lexicalization
- Supervised learning only

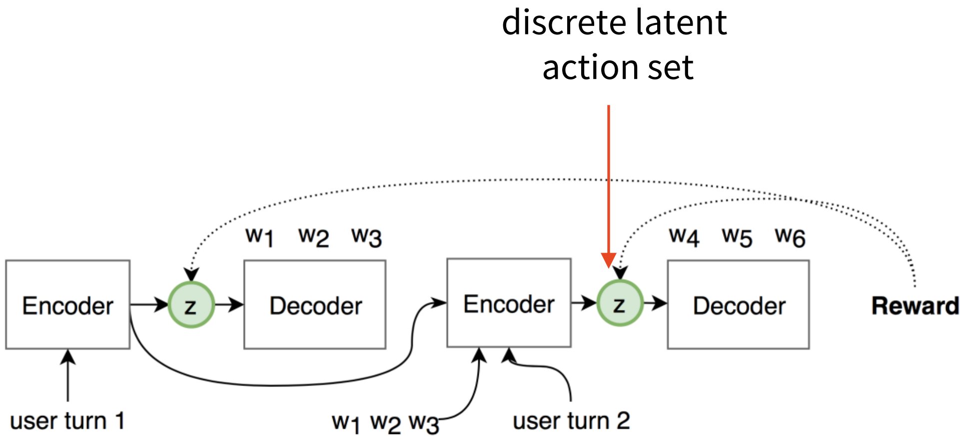


(Chen et al., 2019)
https://www.aclweb.org/anthology/P19-1360

Latent Action RL

(Zhao et al., 2019)
https://www.aclweb.org/anthology/N19-1123

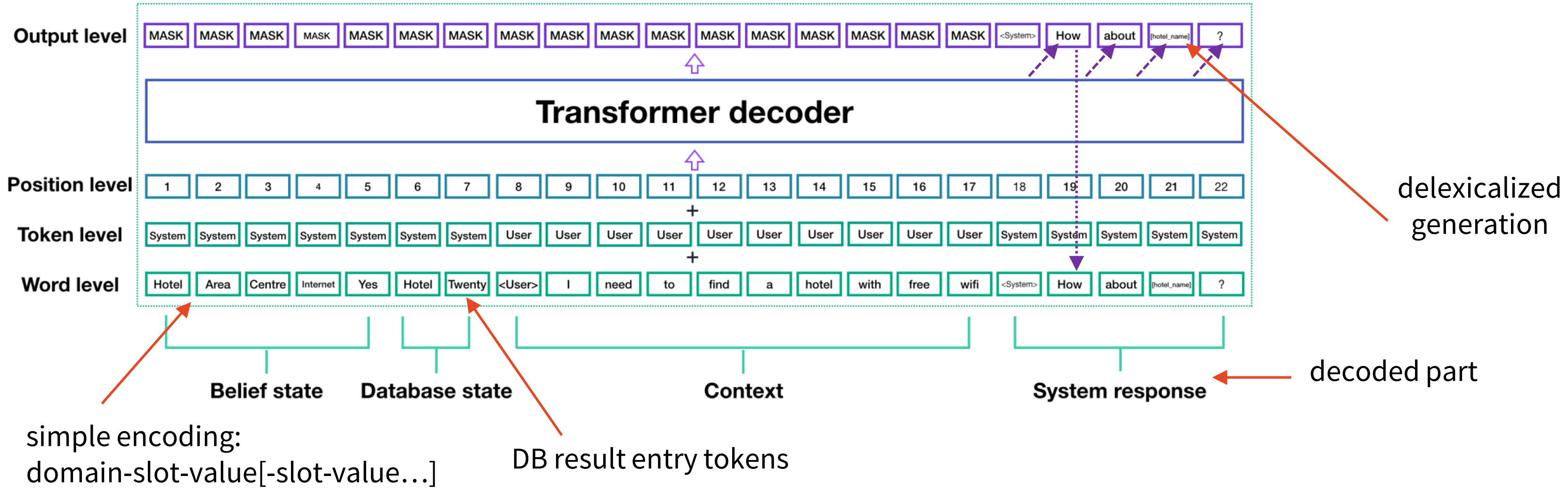
- Making system actions latent, learning them implicitly
- Like a VAE, but **discrete latent space** here (M k-way variables)
  - using Gumbel-Softmax trick for backpropagation
  - using Full ELBO (KL vs. prior network)
    or "Lite ELBO" (KL vs. uniform $1/k$)
- RL over latent actions, not words
  - avoids producing disfluent language
  - "fake RL" based on supervised data
    - generate outputs, but use original contexts from a dialogue from training data
    - success & RL updates based on generated responses
  - on par with Structured Fusion Nets (slightly higher success, lower BLEU)
- again, ignores DB & belief tracking



“Hello, it’s GPT-2 – How can I help?”

(Budzianowski & Vulić, 2019)
https://www.aclweb.org/anthology/D19-5602

- Simple adaptation of the GPT pretrained LM
  - system/user embeddings
    - added to Transformer positional embs. & word embs.
  - training to generate as well as classify utterances (good vs. random)
    - all supervised
- Again, no DB & belief tracking
  - using gold-standard belief & DB, no way of updating belief



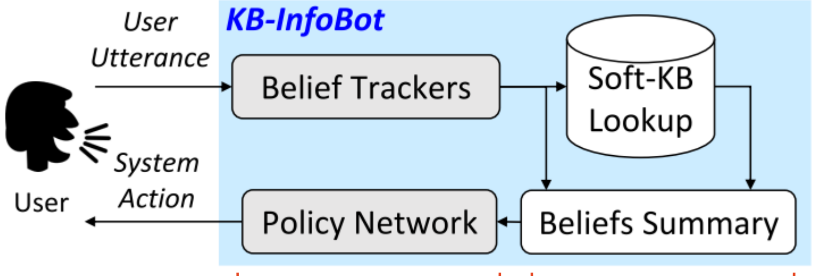
Soft DB Lookups

(Dinghra et al., 2017)
https://www.aclweb.org/anthology/P17-1045

- incorporating NLU/tracker uncertainty into DB results
- making the system fully differentiable
  - but less interpretable
- DB output = distribution over all items
  - plain MLE estimation: $p(\text{row } i) = \prod_{\text{slots } j} \begin{cases} \frac{p(v=j)}{\#\text{ of } v'\text{s in table}} & \text{if } j \text{ specified \& in table} \\ 1/\#\text{ rows (uniform)} & \text{otherwise} \end{cases}$
  - not trained, based directly on tracker

as given by tracker

- NLU/trackers – per-slot GRUs + softmaxes
  - input: counts of n-grams
- policy = GRU + softmax
- trained by RL
  - shown to outperform hard DB on a movie domain



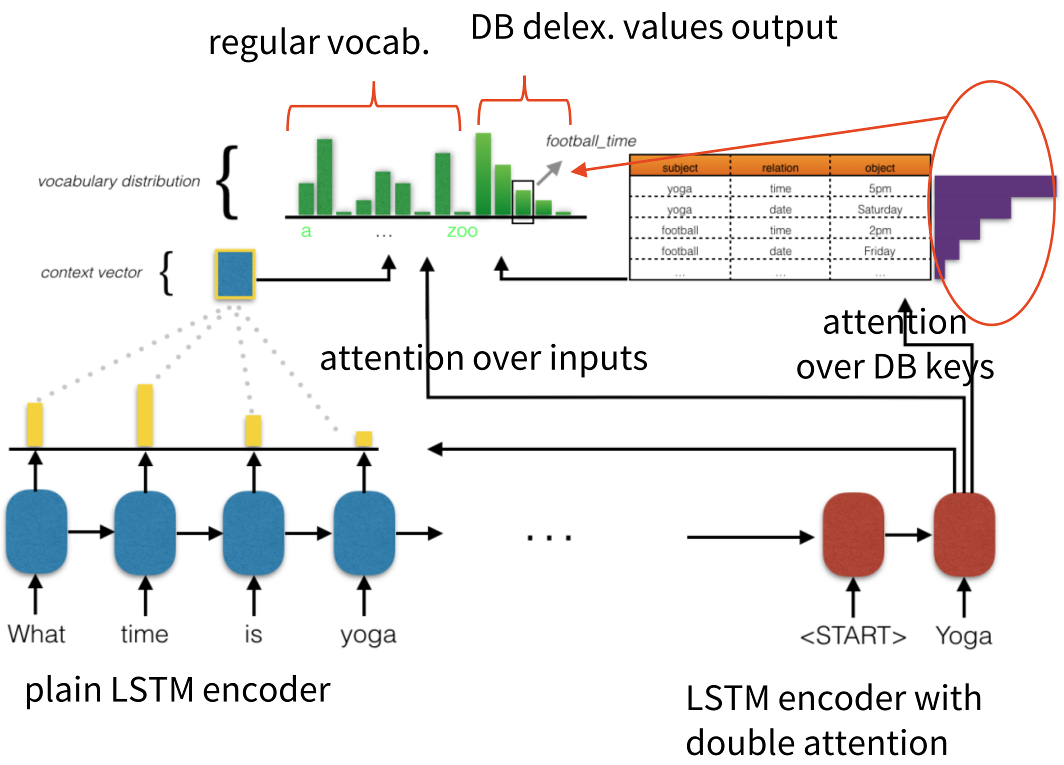
trainable

fixed
(but differentiable)

Key-value Retrieval Nets

(Eric et al., 2017)
https://www.aclweb.org/anthology/W17-5506

- using attention to model DB access
- LSTM encoder, no specific tracker/NLU
- DB in a “key-value” format
  - subject-relation-object (subject-property-value) *dinner-time-8pm*
  - key = subject + relation
    value = subject_relation
    - i.e. delexicalized values

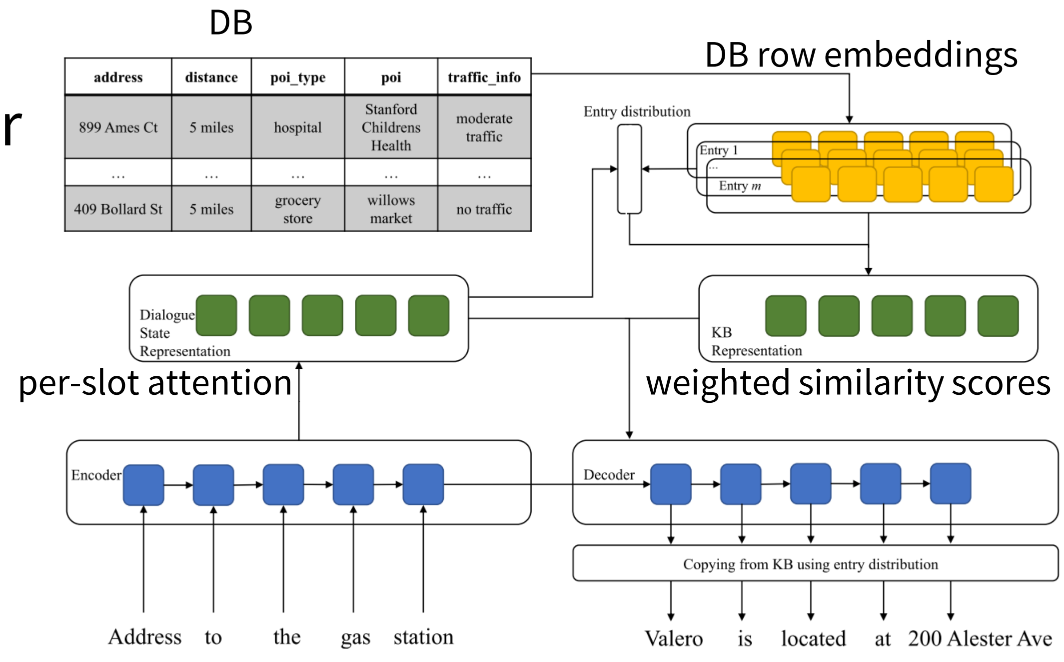


- generator: seq2seq with 2 attentions
  - over inputs (as usual)
  - over <u>keys</u> in the DB – increases generator output probs. of DB <u>values</u>
    - doesn’t change probs. of regular vocabulary
- supervised training, better than seq2seq/copy

DB Table Attention

(Wen et al., 2018)
http://arxiv.org/abs/1806.04441

- Input/State tracking:
  - LSTM encoder over whole history
  - slot states = per-slot attention over encoder
- DB representation:
  - **cell embedding** = column/slot emb.
    concat → & value emb. + linear + tanh
  - **row similarity** with dialogue state:
    $\sum_{\text{slots}} \text{cell emb} \cdot \text{slot state}$
  - **info matrix**: softmax-weighted sum of row similarities
  - **memory**: weights · (slot states & info matrix)
- Response decoder: seq2seq + “copy”
  - with attentions over input & memory
  - copying: choosing to generate a slot & filling in value based on info matrix



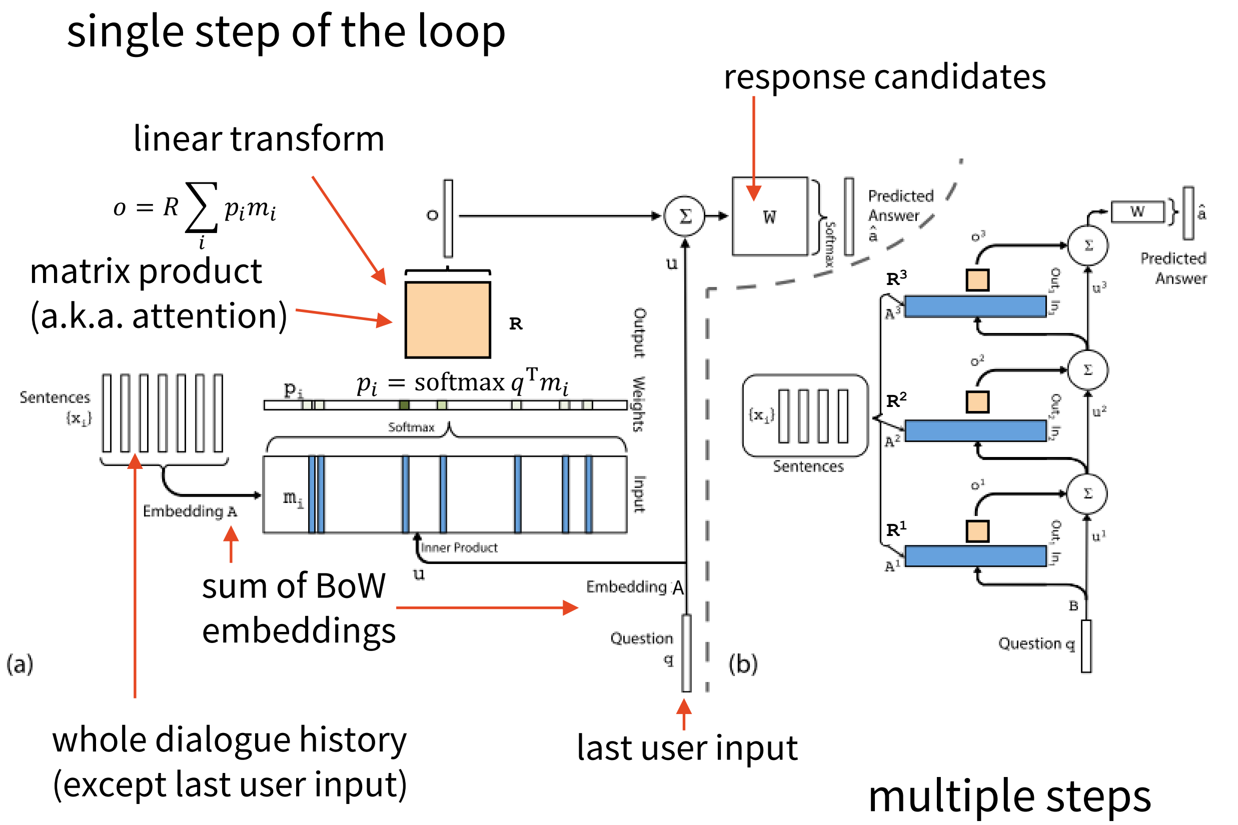
decoder with double attention

Memory networks

(Sukhbaatar et al., 2015)
http://arxiv.org/abs/1503.08895
(Bordes et al., 2017)
http://arxiv.org/abs/1605.07683

- not a full dialogue model,
  just ranker of candidate replies
- no explicit modules
- based on attention over history
  - sum of bag-of-words embeddings
    - added features (user/system, turn no.)
  - weighted match against
    last user input (dot + softmax)
  - linear transformation to produce
    next-level input
- last input matched (dot + softmax)
  against a pool of possible responses

loop
a few
times



(Madotto et al., 2018)
https://www.aclweb.org/anthology/P18-1136

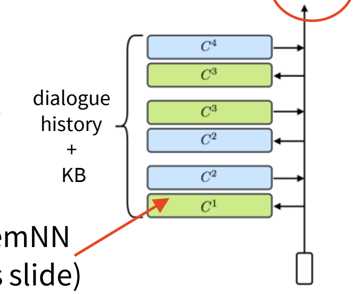
Mem2Seq: memory nets & pointer-generator

- “standard” MemNN encoder:
  - special memory:
    - token-level dialogue history
      (whole history concatenated, no hierarchy)
      - with added turn numbers & user/system flags
    - DB tuples (sums of subject-relation-object)
    - “sentinel” (special token)
- decoder: MemNN over GRU
  - GRU state is MemNN initial query
  - last level attention is copy pointer
  - if copy pointer points at sentinel, generate from vocabulary
    - copies whenever it can
  - vocabulary distribution comes from 1st level of memory + GRU state
    - linear transform + softmax

encoder

state embedding
(1st decoder GRU input)

token-level
dialogue history



standard MemNN
(see previous slide)

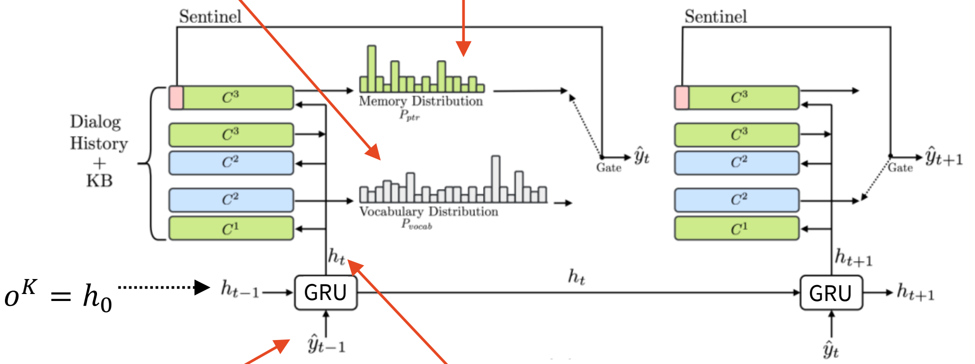
last user input

decoder (word level)

vocab softmax generated
from 1st memory hop
$P_{\text{vocab}}(\hat{y}_t) = \text{softmax}(W_1[h_t, o^1])$

pointer softmax is
last memory level attention
$P_{ptr} = p_t^K$

only if P_{ptr} points
at sentinel,
P_{vocab} is used



previous
generated
word

GRU hidden state is the
initial MemNN query

Mem2Seq

attention visualization

DB

values
(these get output)

subject-relation-object
(this gets embedded)

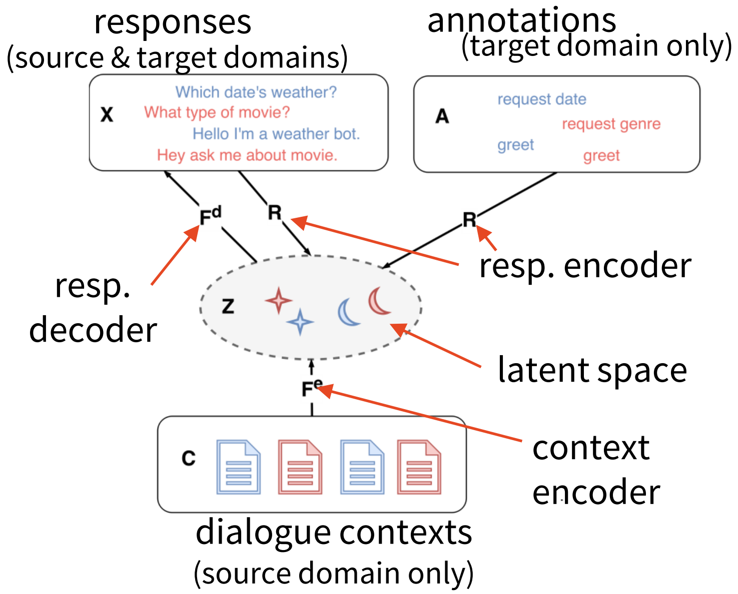
dialogue
history

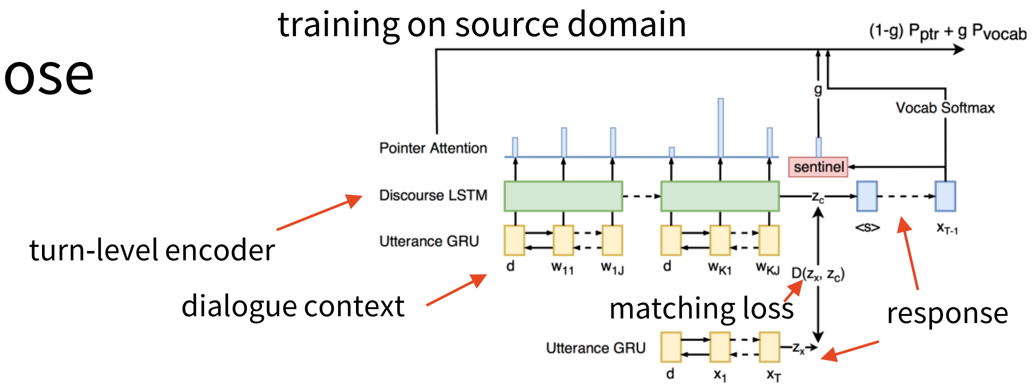
sentinel
“don’t copy, generate”

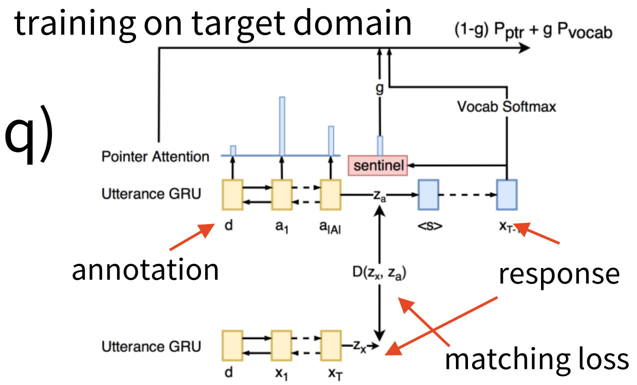
(Zhao & Eskenazi, 2018) http://aclweb.org/anthology/W18-5001

Few-shot dialogue generation

- Domain transfer:
  - source domain training dialogues
  - target domain “seed responses” with annotation
- encoding all into latent space
  - keeping response & annotation encoding close
  - keeping context & response encoding close
  - decoder loss + matching loss
- encoder: HRE (hierarchical RNN)
- decoder: copy RNN (with sentinel)
  - “copy unless attention points to sentinel” (see Mem2Seq)
- DB queries & results treated as responses/inputs
  - DB & user part of environment



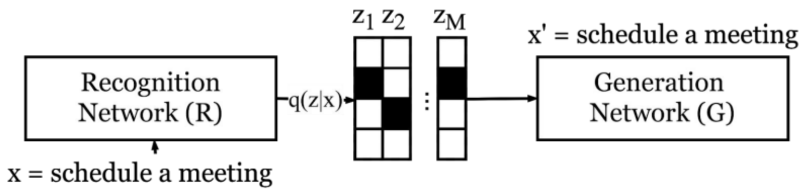


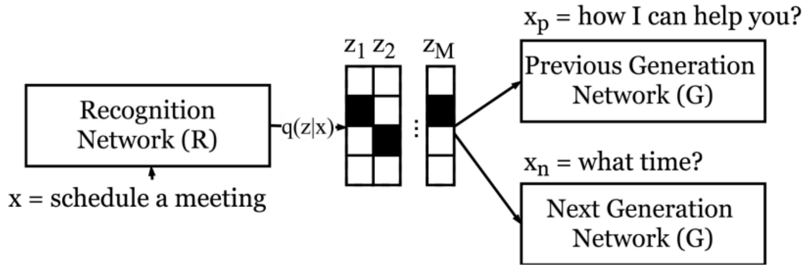


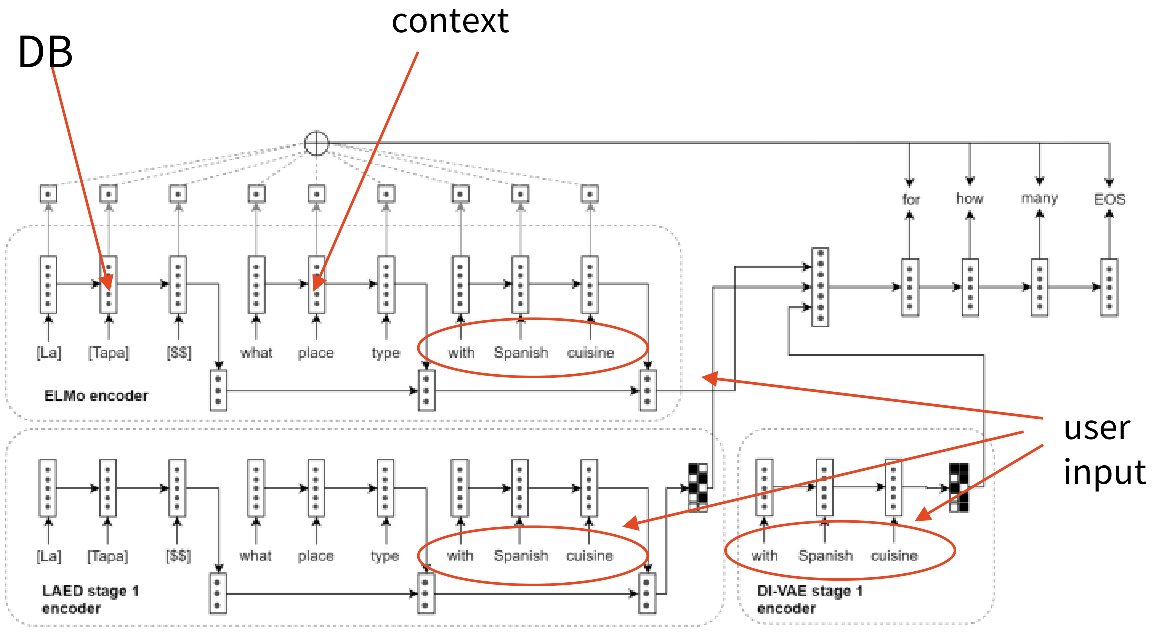
Few-shot & Latent Actions

(Zhao et al., 2018) http://aclweb.org/anthology/P18-1101
https://www.cs.cmu.edu/~tianchez/data/ACL2018-talk.pdf
(Shalyminov et al., 2019) http://arxiv.org/abs/1910.01302

- Latent discrete encoder-decoder
  - discrete VAE for dialogue turns
  - discrete Variational Skip Thought
    - predicting next turn
  - trained jointly
- Full model:
  - LAED to predict next action
  - DI-VAE for user input representation
  - HRED with ELMo
  - KVRET-like DB representation
    - DB is treated as part of context
  - decoder: same as previous
    - copy with sentinel
  - uses NER/entity linking instead of handcrafted annotations







Summary

- RL for end-to-end systems helps if it's not on token level
  - RL over latent system actions (embeddings / discrete)
- Pretrained LMs can work as end-to-end DS
- Soft DB lookups – making the whole system differentiable
  - "transparent" (directly based on tracker)
  - attention/memory nets (multi-hop attention)
- Few-shot: lot of autoencoding

Thanks

Contact us:

odusek@ufal.mff.cuni.cz
hudecek@ufal.mff.cuni.cz
(or on Slack)

Get these slides here:

http://ufal.cz/npfl099

References/Inspiration/Further:

- Gao et al. (2019): Neural Approaches to Conversational AI: https://arxiv.org/abs/1809.08267
- Serban et al. (2018): A Survey of Available Corpora For Building Data-Driven Dialogue Systems: http://dad.uni-bielefeld.de/index.php/dad/article/view/3690