Better Supervision for End-to-end Neural Dialogue Systems

Ondřej Dušek

VGS Invited Talks @ FIT

1.12.2021

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Credits

All this is mainly work of my students, namely

Vojtěch Hudeček

Jonáš Kulhánek

Tomáš Nekvinda

2nd part started during Vojta's internship with

Zhou Yu

Outline

1. AuGPT: GPT-2 for task-oriented dialogue

- background: multi-domain dialogue
- GPT-2 + extensions (model & data)
- results, analysis

2. Weakly supervised slot discovery

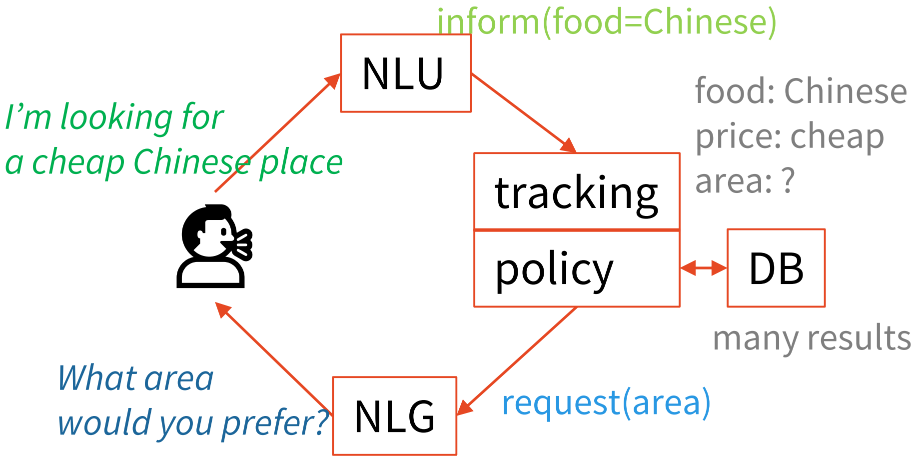
- background: dialogue annotation
- filtering/refining automatic annotation for use in dialogue
- application in an end-to-end system

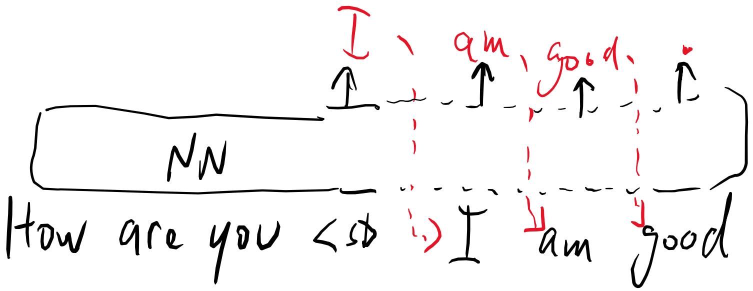
3. Dialogue evaluation

- MultiWOZ evaluation inconsistency
- how to fix it

Part 1 – AuGPT: Background

- trend: **end-to-end neural dialogue systems**
  - single neural network as a whole dialogue system:
    - language understanding
    - belief state tracking
    - dialogue policy (dialogue management/action selection)
    - response generation (word-by-word)
  - database access external
  - typically text-only (as are all suitable training datasets)
- problems:
  - **needs a lot of data** & annotation
    - hard to get by
    - noisy
  - neural nets **hallucinate**
  - **repetitive**/dull outputs





Multi-domain dialogue (MultiWOZ, DSTC9)

- MultiWOZ: benchmark for task-oriented dialogue in multiple domains
  - hotels, restaurants, attractions, trains, taxi, police, hospital
  - domains are connected (e.g. taxi from hotel to attraction)
  - 10k dialogues, extensive annotation (but noisy!)

1.
user: I am looking for a train from Cambridge to London Kinks Cross.
state: *{train {departure = cambridge, destination = london kings cross}}*
DB: *{train (70) {…}}*
[count] [departure] [destination]
system: There are 70 trains departing from Cambridge to London Kings Cross. What day would you like to travel?

2.
user: I would like to leave on Saturday after 18:45.
state: *{train {day = saturday, departure = cambridge, destination = london kings cross, leave at = 18:45}}*
DB: *{train (3) {arrive by = 19:51,21:51,23:51; id = TR0427,TR0925,TR4898; leave at = 19:00,21:00,23:00; … }}*
[id] [leave_at] [arrive_by]
system: TR0427 leaves at 19:00 and arrives by 19:51 . Would you like to book this train?

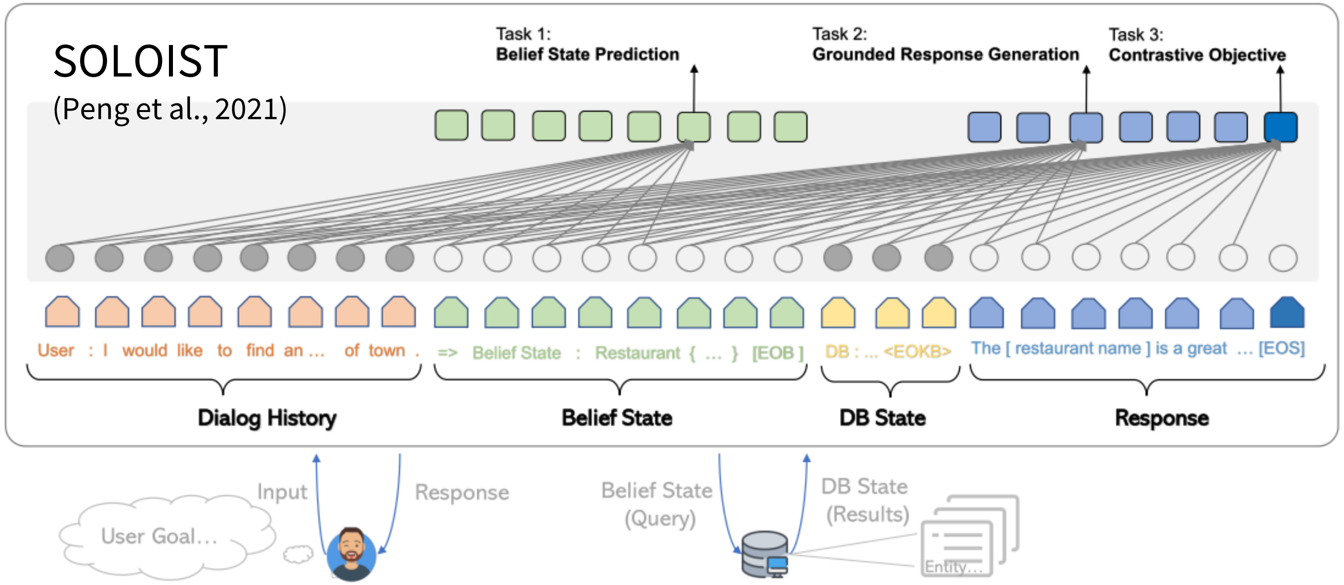
3.
user: Yes, I would like to book it for eight people.
[reference]
system: I have booked it for you. Your reference number is 00000057. Is there anything else I can help you with?

4.
user: I am also looking for an expensive restaurant in the centre.
belief: *{restaurant {area = centre, price range = expensive} train {...}}*
DB: *{restaurant (33) {area = centre (33); name=Curry Garden, ...; ...}, ...}*
[count] [price_range] [area]
system: There are 33 expensive restaurants in the centre. Is there a particular type of food you would like?

End-to-end Neural Dialogue with GPT-2

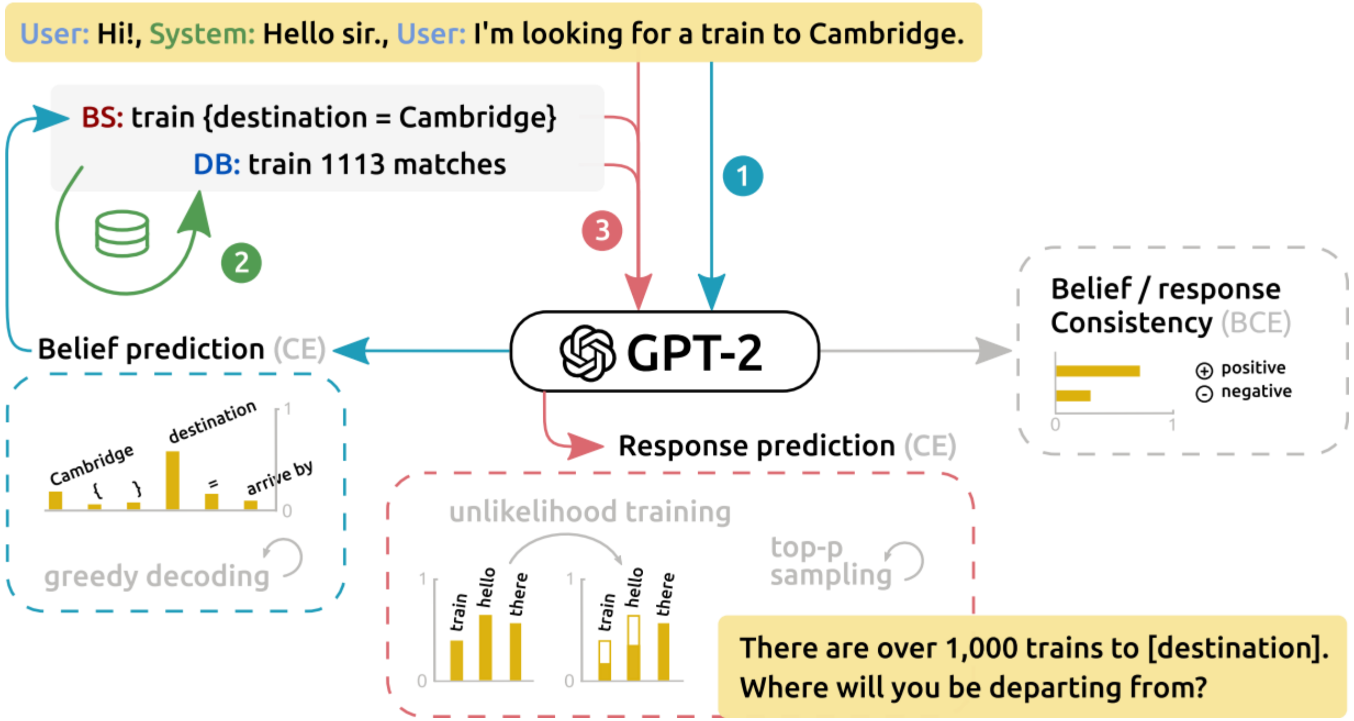
- Multiple recent DSs are based on GPT-2 (SOLOIST, UBAR, SimpleTOD, NeuralPipeline)
  - huge NN pretrained on next-word generation, helps with fluency
- Everything is sequence generation
  - dialogue context, belief state, database outputs represented as sequences
- Multi-step operation:
  1) prompt with context & decode belief state
  2) query DB (external)
  3) prompt with DB output & decode response

SOLOIST
(Peng et al., 2021)



AuGPT architecture

- Same principle, multiple improvements
  - based on SOLOIST
- Operation:
  1) context → belief state
     - greedy decoding
     - text-like belief state
  2) belief state → DB
     - text-like DB results
  3) DB → response
     - top-p sampling (diversity)
     - delexicalized (slot placeholders)
- Training:
  - belief/response prediction + consistency (Y/N)



Consistency task

- **Additional training task** – generating & classifying at the same time
  - additional classification layer on top of last decoder step logits
  - incurs additional loss, added to generation loss
- Aim: **robustness** – detecting problems
  - **½ data artificially corrupted** – state or target response don't fit context
  - SOLOIST: corrupted state sampled randomly
  - **AuGPT**: corrupted state sampled from the **same domain – harder!**

context	state	**response**	**consistent?**	
i want a cheap italian restaurant	*{ price range = cheap , food = Italian }*	ok which area ?	✓	
i want a cheap Italian restaurant	*{ price range = cheap , food = Italian }*	thanks, goodbye !	✗	bad response
i want a cheap italian restaurant	*{ destination = Cambridge , leave at = 19:00 }*	ok which area ?	✗	bad state
i want a cheap italian restaurant	*{ area = north , food = Chinese }*	ok which area ?	✗	bad state (same domain)

SOLOIST

new in AuGPT

AuGPT improvements

- **Better consistency** auxiliary training task (↑)
- **Data augmentation** via backtranslation (en → xx → en)
  - MT between English and 40 languages from the ELITR project (https://elitr.eu/)
  - we chose 10 best languages
  - user inputs chosen at random from **original & 10 backtranslated texts**
- **Data cleaning**
  - checking consistency of user goal with database
  - ~30% MultiWOZ data discarded
- **Unlikelihood loss** for output diversity
  - repeated tokens are penalized
- **Sampling** for output diversity

Results

- **Corpus-based** evaluation: competitive
  - 1-turn – uses gold contexts
- **Simulator**: much better
  - runs whole dialogue
  - good 1-turn ≠ good over whole dialogue!
- DSTC9 competition – **humans**
  - beats the baseline (4 out of 10)
  - 3rd overall, 1st if DB consistency ignored
  - shortest dialogues needed
- **Ablation**
  - simulator + corpus
  - confirms our new features (unlikelihood & data cleaning – sim only)

	succ	**inf**	**BLEU**
SOLOIST	85.5	72.9	16.5
DAMD	76.3	60.4	16.6
LAVA	91.8	81.8	12.0
AuGPT	83.1	70.1	17.2

Corpus-based MultiWOZ 2.0

	cpl	**succ**	**inf F1**	**book**
DAMD	39.5	34.3	56.3	51.4
AuGPT	89.4	60.1	70.1	85.7

ConvLab 2 simulator

	succ+DB	**succ-DB**	**# turns**
baseline	56.8	82.4	18.5
winner	70.2	79.4	18.5
AuGPT	62.0	82.6	17.1

DSTC9

Error analysis

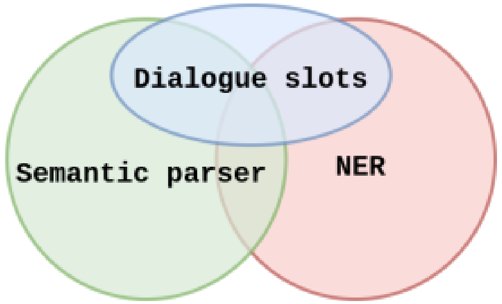
- We talked to the system & recorded errors
- 130 dialogues, goals from test set
- 50 erroneous dialogues, 17 unsuccessful
  - success rate ~87%
  - **most errors recoverable**
  - **correct behavior** in many non-trivial cases
- Errors categorized
  - hallucinated values (21) – non-grounded slots
  - wrong lexicalization (6) – repeated values
  - missing information (5) – over-eager booking
  - ignored input (5) – asks for the same thing repeatedly
  - …

AuGPT bottom line

- GPT-2 + augmentation + cleaning + consistency → really good results
  - especially if you go beyond evaluating on 1 turn only
- Still not perfect
  - main problem: grounding
- Also, needs a lot of annotated data
  - MultiWOZ is large & was expensive to collect

Part 2 – Slot discovery: Motivation

- Dialogue annotation is expensive, we need a lot of it
  - it’s domain/task-specific – does not translate across domains
- There’s plenty of generic automatic annotation tools
  - named entities
  - semantic roles
  - ….



Dialogue slots: I am looking for a cheap restaurant in Georgetown.

Semantic parser: I am looking for a cheap restaurant in Georgetown.

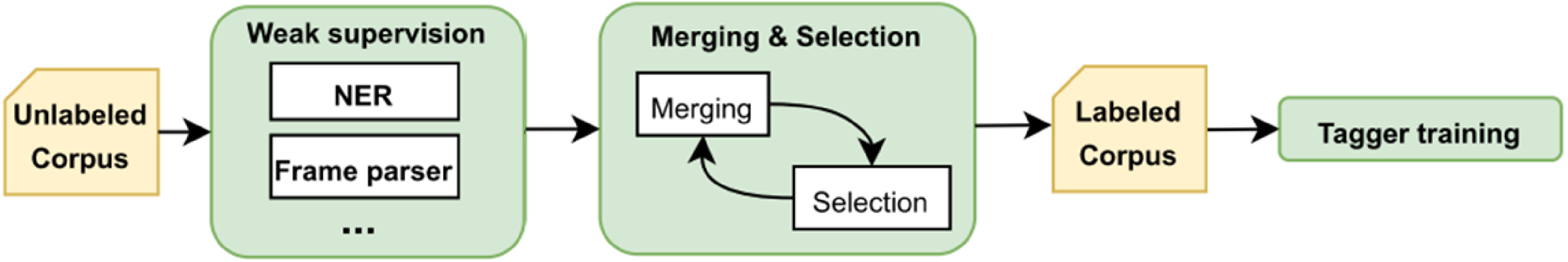
NER: I am looking for a cheap restaurant in Georgetown.

- Can we get domain-relevant dialogue slots from generic annotation?
  1) weak supervision from generic annotation
  2) filtering, refinement
- How well would they work in an end-to-end system?

Weakly supervised slot discovery

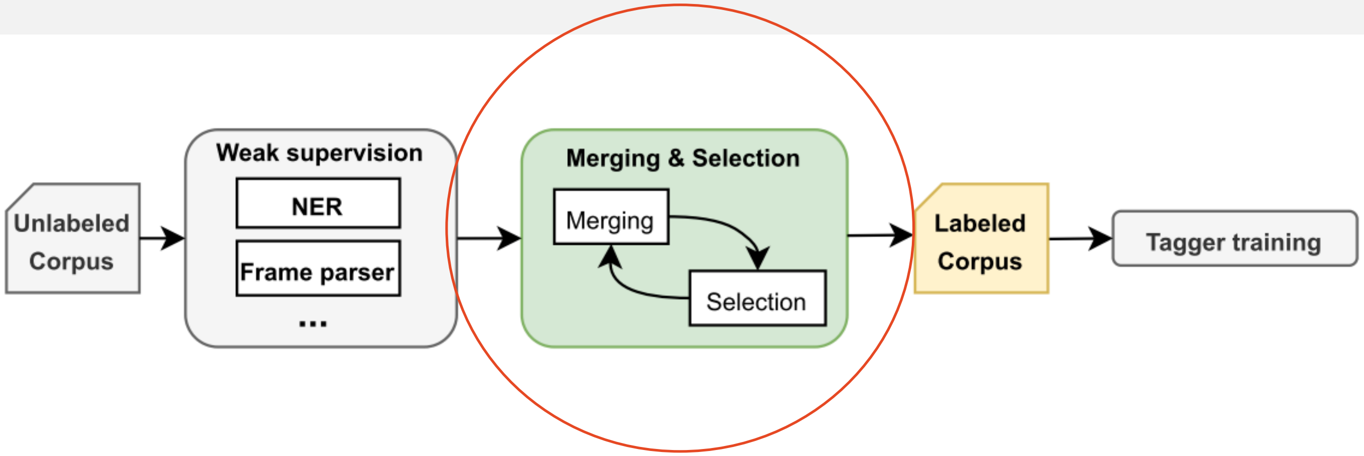
Overall approach:

1) annotate by generic parsers/taggers
   - use high label granularity (mainly frame semantics)
2) merge labels & select relevant only
   - iterative process
3) using the final labels, train a standalone tagger
   - no need to do use the original parsers/taggers anymore
   - can actually improve over them



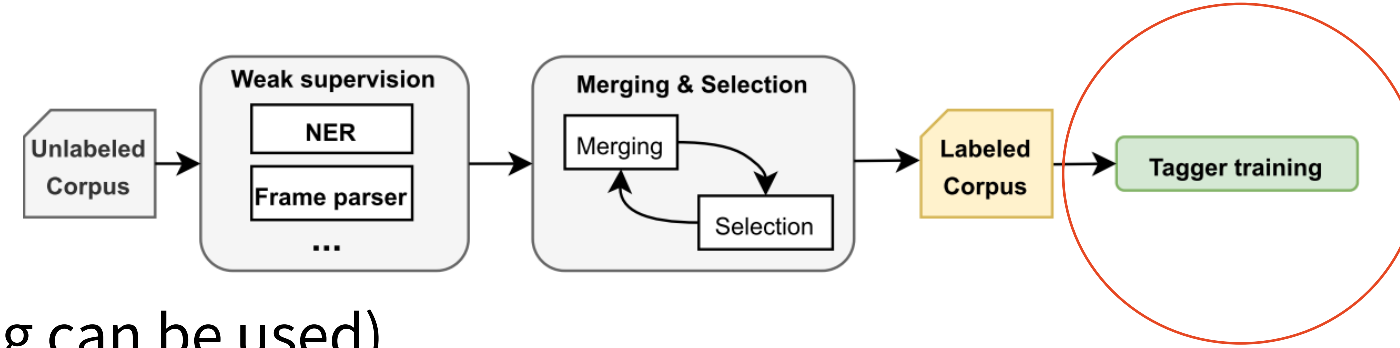
Slot merging & selection

- Assuming slot candidates
  - works with slot candidates, not individual occurrences (fillers)



- In each iteration:
  1) compute slot candidate embeddings
     - average FastText of all slot fillers
  2) **merge** candidates based on similarity
     - embedding cosine similarity + occurrence in same contexts
     - > threshold → merge
  3) **select**: rank candidates, then remove low-ranked ones
     - split into clusters according to contexts
     - ranking: frequency + coherence (embedding similarity) + TextRank
     - $> \alpha \cdot$ cluster mean → keep
- Repeat until convergence (no changes)

Standalone tagger



- **Refined annotation** used **to train** a standalone **tagger**
  - LSTM sequence tagger (but anything can be used)
  - labels: none / slot-0, slot-1 … (from our annotation)
- **No need for the original parsers** anymore
  - ready to use on new data
- Labels are sparse → use **lower confidence threshold** to improve recall
  - p(none) > 0.5 & p(slot-X) > 0.2 → tag as slot-X

Example

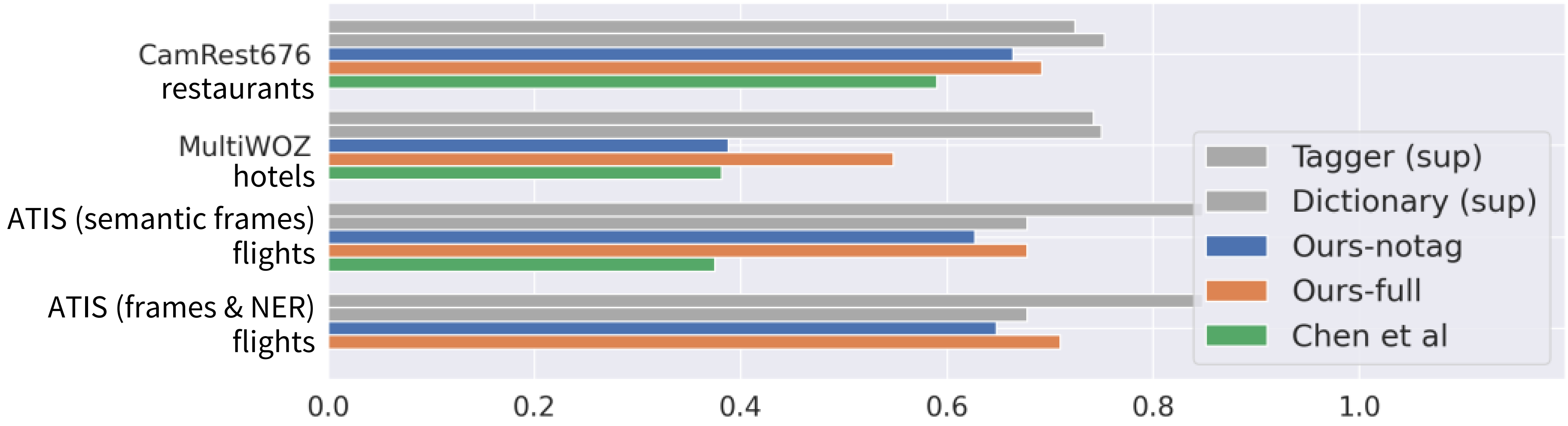
- Irrelevant slots removed
  - domain is all about restaurants, all have food
- Meaningful slots kept
  - the same slot points to similar entities
- Tagger generalizes beyond original annotation
  - “Afghan” recognized

User input 1:	I would like an	expensive	restaurant	that serves	Afghan	food.
Original annotation:		Expensiveness	Locale			
Our annotation:		**slot-0**			**slot-1**	

User input 2:	How about	Asian	oriental	food.
Original annotation:		Origin		Food
Our annotation:		**slot-1**		

Results

- Slot F1 on multiple datasets
  - compared to supervised methods (LSTM tager/dictionary)
    + previous weakly supervised method (Chen et al.)
  - with & without standalone tagger (full/notag)
- Worse than supervised, but not so far, better than previous
  - ATIS: NER helps on top of semantic frames
  - limited by the underlying annotation, some errors in merging



Use in end-to-end systems

- LSTM-based 2-stage copy network (Jin et al., 2018)
  - encode context → decode state → query DB → decode response
  - basically AuGPT minus GPT, but with LSTMs & explicit copying
  - unsupervised version: reconstruction loss, state as memory
- Ours better than unsupervised, better than semantic frames alone
  - especially in whole dialogues (correct entity found)
  - label quality is important: unfiltered semantic frames → slot F1 drop

	slot F1	joint goal accuracy	entity match rate
supervised	96.7	89.7	86.9
unsupervised	71.9	38.5	1.9
semantic frames	70.9	33.5	26.9
Ours-full	75.6	46.5	36.8

Results on CamRest676 (restaurants)

Slot discovery bottom line

- Domain-relevant slots can be found using generic annotation tools
- Resulting annotation is noisy but still helpful
- Partial manual annotation / better source annotation may help
  - completely unsupervised – keyword extraction
- So far, annotation on single sentences only
  - context may help

Part 3: MultiWOZ benchmark caveats

- MultiWOZ evaluations use different implementations
  - but the numbers are compared across papers!
- **Preprocessing** – delexicalization (slot placeholders)
  - based on string matching
- **Postprocessing** – lexicalization
  - evaluation goes over placeholders
- **Database** – value match
  - normalization is needed
  - booking availability is random
- **Metrics**: BLEU, inform & success
  - BLEU tokenization
  - how to match imperfect entity overlap (goal vs. found)?
  - no output diversity metrics

cafe jello gallery has a free entrance fee.
[address] has a **[entrancefee]** entrance fee.

4pm = 16:00
the botanical gardens at cambridge university
= cambridge university botanical gardens

Delexicalization example

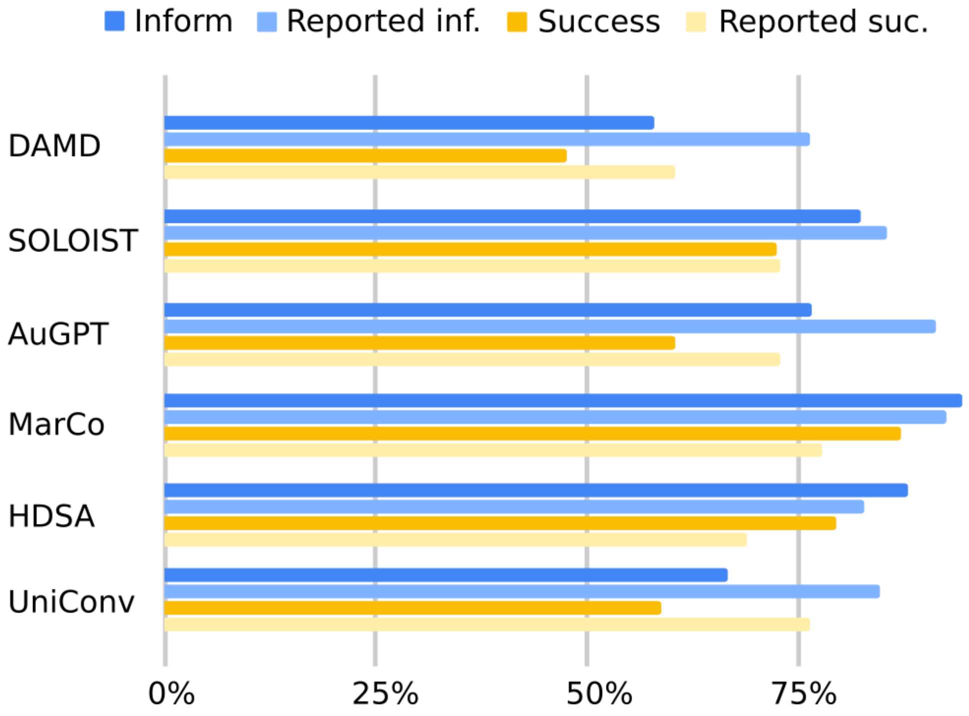
Original	**Cafe jello gallery** has a **free** entrance fee. The address is **cafe jello gallery, 13 magdalene street** and the postcode is **cb30af**. Can i help you with anything else?
MultiWOZ 2.2	**[address]** has a **[entrancefee]** entrance fee. The address is **[name]**, **[address]** and the post code is **[postcode]**. Can I help you with anything else?
DAMD	**[value_name]** has a **[value_price]** entrance fee. The address is cafe jello gallery, **[value_address]** and the postcode is **[value_postcode]**. Can i help you with anything else?
HDSA	**[attraction_name]** has a free entrance fee. The address is **[attraction_address]** and the post code is **[attraction_postcode]**. Can I help you with anything else?
AuGPT	**[address]** has a free entrance fee. The address is cafe jello gallery, **[address]** and the post code is **[postcode]**. Can I help you with anything else?
UniConv	**[attraction_name]** has a **[attraction_pricerange]** entrance fee. The address is **[attraction_name]**, 13 **[attraction_address]** and the post code is **[attraction_postcode]**. Can i help you with anything else?

What we did

- Re-evaluated outputs of 13 end-to-end/policy dialogue systems
- **Unified delexicalization** styles
- **Unified BLEU**
  - same references
  - SacreBLEU (standardized implementation)
- **Evaluated Inform & success** under identical conditions
  - same DB fuzzy matching
  - same set overlap criteria
- **Evaluated diversity**
  - distinct n-grams
  - entropy
  - MSTTR

Results

- Scores were not comparable
  - now they are, with our standardization
- **BLEU:** delexicalization & tokenization make up to ±2% BLEU
- **Inform & success**: up to 20%↑ on both rates
- **Diversity**:
  - generally low
  - e.g. $^{1}/_{10}$- $^{1}/_{2}$ human vocabulary size in system outputs
  - especially in models trained with RL



Conclusions

- **AuGPT:** Pretrained models work great for task-oriented dialogue
  - problem: **consistency**
  - auxiliary tasks help, to a point
  - problem: **output diversity**
  - data augmentation & unlikelihood helps
- **Slot discovery** is possible from generic annotation
  - annotation needs to be filtered & clustered
  - improves an end-to-end dialogue system
- **Data & evaluation** are as important as models
  - clean data – as much as possible
  - standardized evaluation is needed (ours is now standard for MultiWOZ)

Thanks

References & code:

- AuGPT:
  - Kulhánek et al., NLP4ConvAI 2021 – http://arxiv.org/abs/2102.05126
  - https://github.com/ufal/augpt/
- Slot discovery:
  - Hudeček et al., ACL 2021 – https://aclanthology.org/2021.acl-long.189
  - https://github.com/vojtsek/joint-induction
- MultiWOZ evaluation:
  - Nekvinda & Dušek, GEM 2021 – https://aclanthology.org/2021.gem-1.4
  - https://github.com/Tomiinek/MultiWOZ_Evaluation

Contact: {odusek,hudecek,nekvinda}@ufal.mff.cuni.cz, jonas.kulhanek@cvut.cz
http://ufal.cz/ondrej-dusek

These slides: https://bit.ly/ds-supervision