



Universität Regensburg



Ostbayerische
Technische Hochschule
Amberg-Weiden

Counterfactual dialogue mixing as data augmentation for task- oriented dialogue systems

Sebastian Steindl, Ulrich Schäfer, Bernd Ludwig

LREC-COLING 2024

Agenda

- Introduction
- Counterfactuals
- Dialogue Mixing
- Possible Problems due to Mixing
- Evaluation
- Conclusion

Introduction / Background

- Chatbot = Human-Computer Interaction
- **Task-Oriented Dialogue** (TOD) systems (e.g., booking)
- Dialogue Understanding, Policy Planning, Natural Language Generation
- One transformer-based model
- Data collection is **expensive**, no web-scraping.
- → Investigate data augmentation for TOD systems

Counterfactuals

- **Hypothetical** situation, (at least) one element changed
- „**Would my headache have gone away, had I not taken medicine?**“
 - Time passed
- SCM not obtainable

Counterfactual Dialogues

- X: User utterances
- Y: System response
- U_X : exogenous variable
- $f_X(U_X)$: mechanism that creates utterance (cognition)

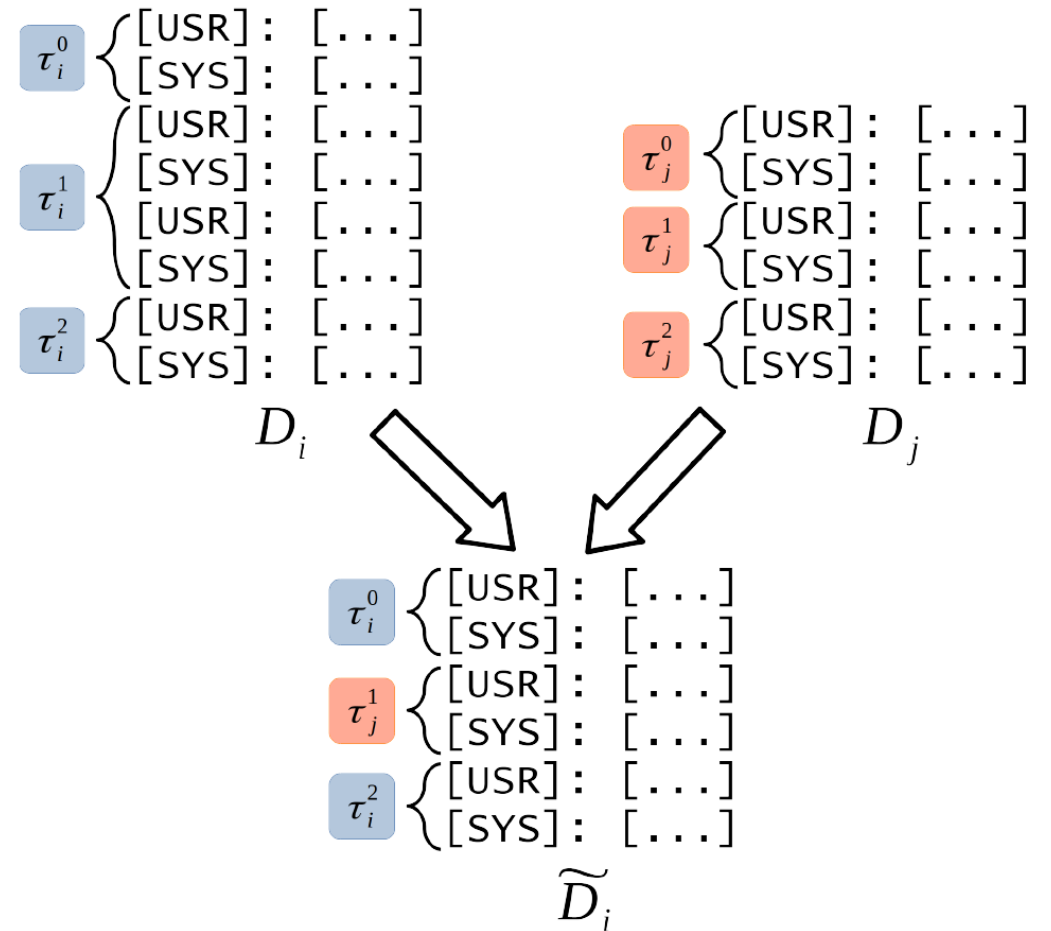
$$X := f_X(U_X)$$

$$Y := f_Y(X) + U_Y$$

- Dialogue mixing procedure
- same topic change \rightarrow fix U_X
- \rightarrow fix the larger goal, change the utterances X

Dialogue Mixing

- MultiWOZ: multiple domains τ_i^n (taxi, hotel, ...)
- equivalent to patches for images
- For a given dialogue
 1. randomly select one domain change
 2. randomly select a dialogue with the same change
 3. mix their utterances
- Multiple rounds possible



Possible Problems due to Mixing

- Dialogue flow: intact due to fixing the domains
- Inconsistencies regarding entities
 - often confined to one domain
 - using delexicalized data (problems remain for binary slots)
 - → mostly unproblematic

Example

[USR]: I am looking for a turkish restaurant

[SYS]: [value_name] is a [value_pricerange] [value_food] restaurant in the [value_area]. Would you like me to book it for you?

[USR]: Yes, please. I need a table for 1 person at 14:00 on monday.

[SYS]: Booking was successfull. The table will be reserved for 15 minutes. Reference number is [value_reference]. Is there anything else I can help you with?



[USR]: I am looking for a moderately priced turkish restaurant

[SYS]: There are [value_choice] [value_pricerange] [value_food] restaurants. Do you have a preference on area of town?

[USR]: No, I don't have a preference. I need a table for 1 at 14:00 on monday.

[SYS]: I have booked you at [value_name]. The table will be reserved for 15 minutes. Reference number is [value_reference].

Evaluation on MultiWOZ

- Approach from Cheng et al. (2022)
- Three models, different optimization policy
- Interactive model-based evaluation with a user simulator
 - Sentence score: quality of a single sentence
 - Session score: quality of the whole dialogue

Metric	Base	Base+CDM
Sentence Score	1.44	1.43
Session Score	0.89	0.92

Table 1: Evaluation of the sentence score and session score Model. For the sentence score lower is better and for session score higher is better.

Results

- For all models, CDM improves the results on 3 out of 4 metrics
- Sentence score is systematically worse with CDM

Model	Inform		Success		Sentence		Session	
	Base	Base+CDM	Base	Base+CDM	Base	Base+CDM	Base	Base+CDM
RL-Succ	95.9	99.1	93.9	99.1	0.799	0.812	0.876	0.957
RL-Sent	94.6	98.6	89.4	97.6	0.746	0.834	0.953	0.959
RL-Sess	95.6	96.7	90.3	96.6	0.73	0.799	0.957	0.962

Table 2: Evaluation of the MTTOD model trained either on the base dataset, or the extended dataset with CDM.

Lower Ressource Setting

- Lower Ressource (LR) setting:
20% training data
- Improvements on task-completion.
- Session score got worse

Metric	LR	LR+CDM
Inform	73.5	88.1
Success	69.5	77.9
Sentence Score	1.01	0.90
Session Score	0.92	0.82

Table 3: Result of a RL-Succ model on the test data, trained in the lower resource setting.

External Evaluation

- Restaurant domain from SGD dataset
- Only language generation improved
- In both cases, task-completion is insufficient

Metric	Base	Base+CDM
Inform	17.93	17.93
Success	17.93	17.93
Sentence	1.19	1.16
Session	0.70	0.84

Table 4: Result of the external evaluation with the RL-Succ model.

Conclusion

- CDM = textual data augmentation through mixing dialogues
- Improvements in normal and LR setting
- Interactive evaluation with a weak User simulator can lead to goal mismatch
- Generalization to external data should be considered in future work



Universität Regensburg



Ostbayerische
Technische Hochschule
Amberg-Weiden

Thank you!

Counterfactual dialogue mixing as data augmentation for task-oriented
dialogue systems

Sebastian Steindl, Ulrich Schäfer, Bernd Ludwig
LREC-COLING 2024