Towards Practical Natural Language Generation (NLG) System in Task-Oriented Dialog

Jinfeng Rao, Facebook AI

Joint work with Anusha, Kartikeya, Michael, Rajen, Anuj

Berkeley NLP Seminar
09/30/2019

Overview

Assistant NLG

Generate natural language from Dialog output

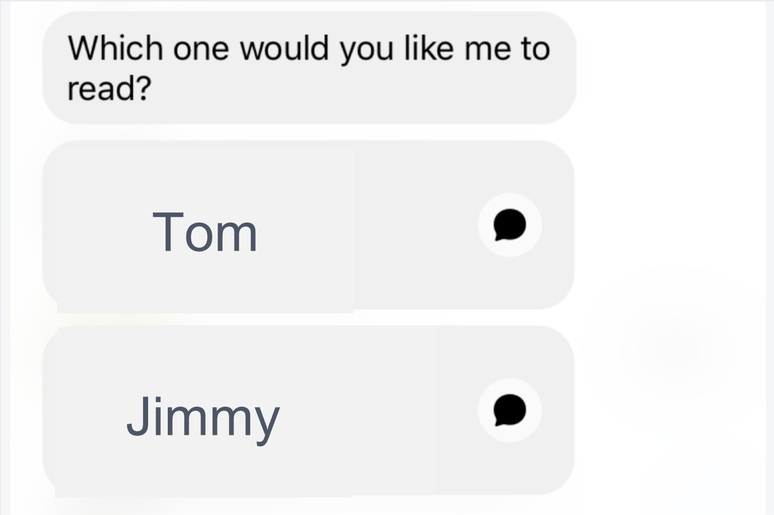
“Do I have any new messages?”

[IN:GET_MESSAGE **Do i have any** [SL:TAG_MESSAGE **new**] **messages**]

goal:agent:messaging_get_message

composer:messaging:inform_read_message

“Your latest chat is with Tom, and you have an unread message from your chat with Jimmy from September 1st. Which one would you like me to read?”

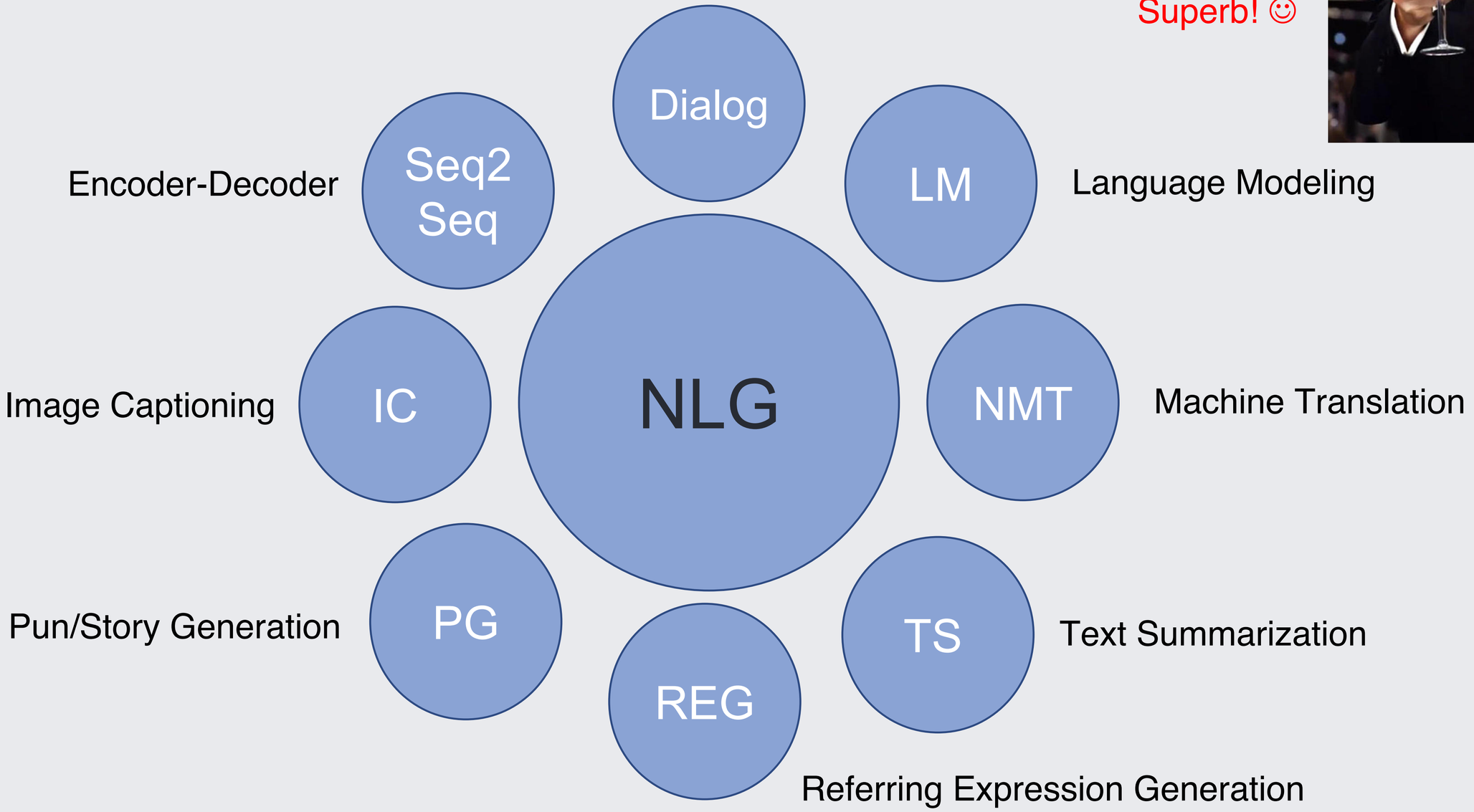


Natural language generation

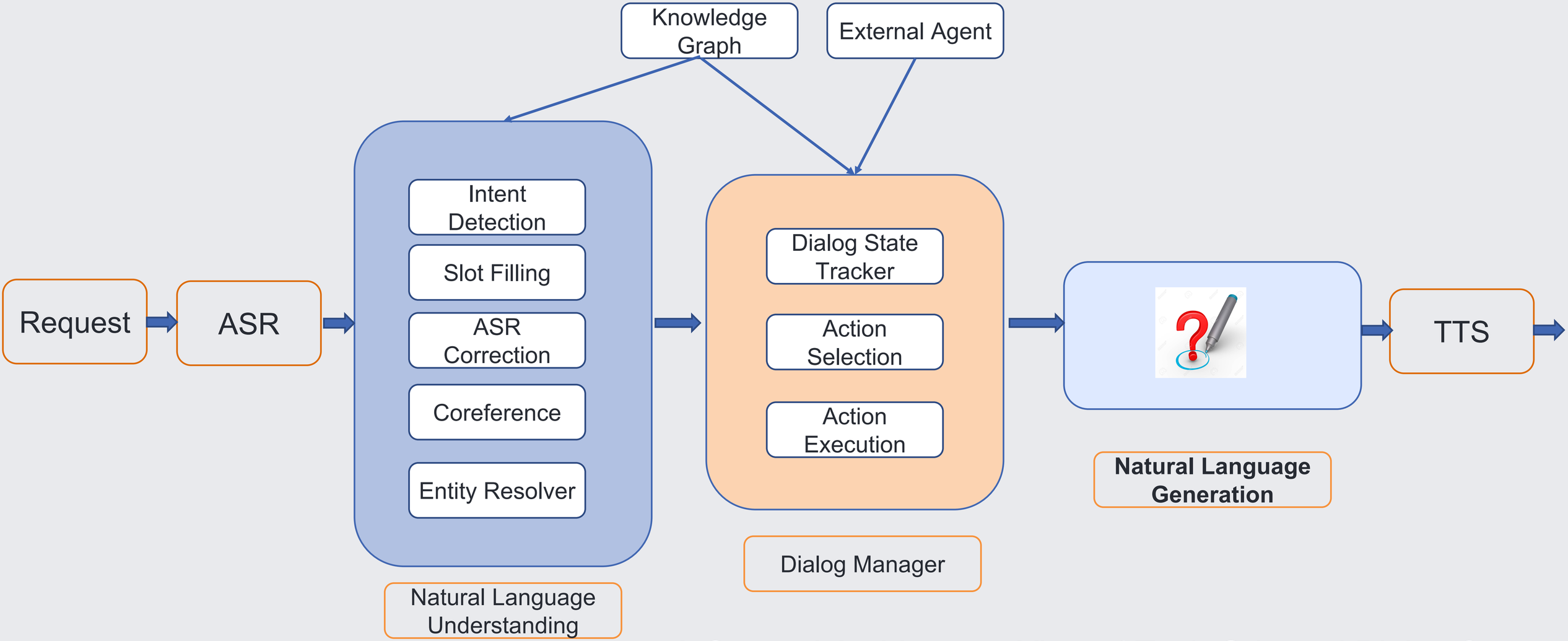
- Naturalness is the primary metric we want to optimize
- We want our system to be natural **in context**
  - Dialog context
  - User context, including memory
  - Multimodal context

Who we think we are

Superb! ☺

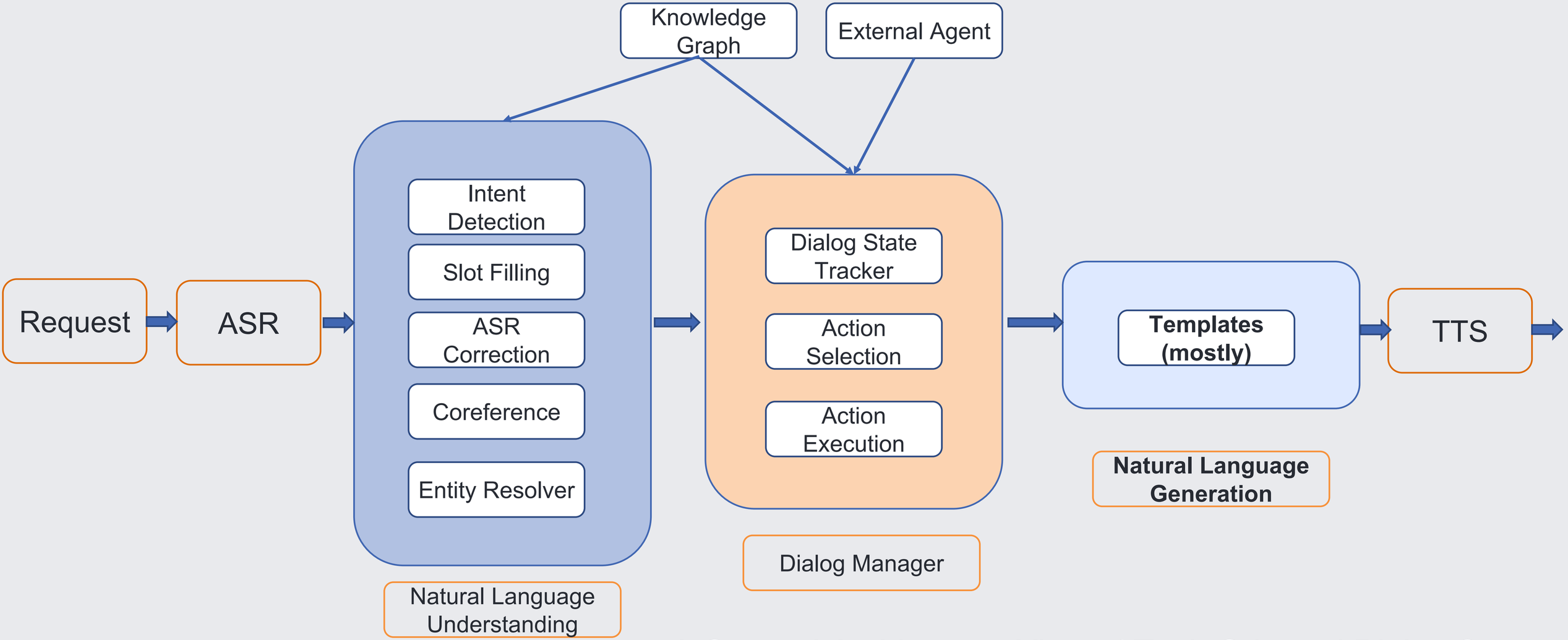


Who we actually are



A General Architecture View of Task-Oriented Dialog System

Who we actually are



A General Architecture View of Task-Oriented Dialog System

Templates

{date_time} in {location}, it'll be {condition} with a high of {temp_high} and a low of {temp_low} F.

Tomorrow in Palo Alto, it'll be cloudy with a high of 64 and a low of 41 F.

March 6th in Golden Gate Bridge, it'll be rain with a high of 64 and a low of 41 F.

- Strengths
  - Simple to bootstrap
  - Robust and interpretable
- Drawbacks
  - Hard to maintain -- lots of templates to handle every case
  - Surface realization rely on argument generators
  - Unable to generalize across domains

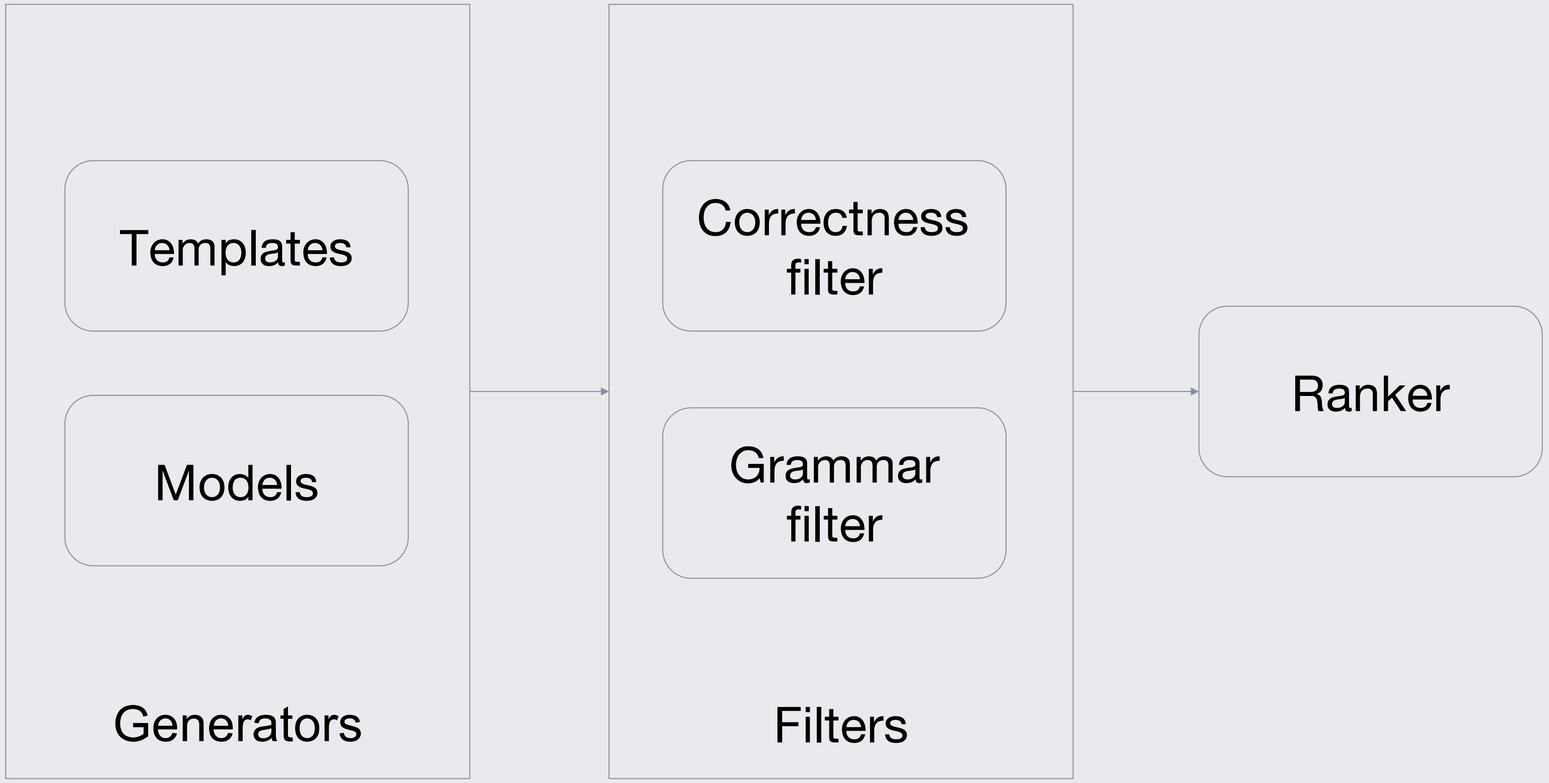
Why Model-based generation?

- Strengths
  - Flexible surface form realization
  - Easy to maintain
  - Reusable across domains
  - Easy to condition on wide feature set (e.g. contextual NLG)
- Drawbacks
  - Adds latency
  - Could generate ungrammatical or incorrect responses

But still… Some Open Questions

- Response Quality
- Controllability
- Evaluation
- Model Robustness
- Cross-domain/language generalizability

Generate-Filter-Rank



Related Papers

- Constrained Decoding with Compositional Meaning Representations for Neural NLG in Task-oriented Dialog, **ACL 2019**

- Tree-to-Sequence Model for Neural NLG in Task-oriented Dialog, **INLG 2019.**

Background

E2E NLG Dataset

~50K (meaning representation, sentence) pairs for the restaurants domain

MR: `name[JJs’ Pub] rating[5 out of 5] familyFriendly[no] eatType[restaurant] near[Crowne Plaza Hotel]`

Sentence: JJ’s pub is not a family friendly restaurant. It has a high customer rating of 5 out of 5. You can find it near the Crowne Plaza Hotel.

E2E Dataset and Shortcomings

- Provided crowdworkers with MR and asked them to write responses

- Lots of diversity in dataset, e.g. argument grouping, contrast, rich language
  - [But no way to control these aspects]

- **However**, models trained on the data lack this diversity
  - e.g. contrast only occurs in 0.4% of model generations

E2E NLG Dataset

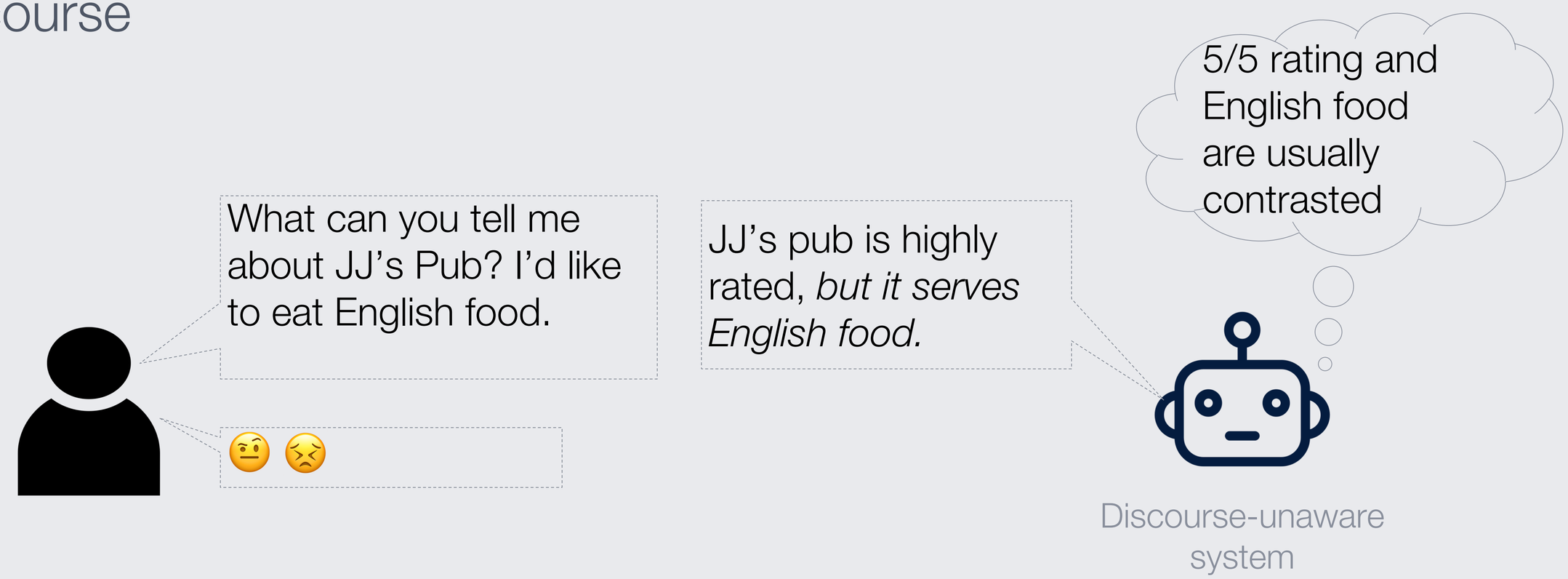
MR: `name[JJs' Pub] rating[5 out of 5] familyFriendly[no] eatType[restaurant] near[Crowne Plaza Hotel]`

***S1**: JJ's pub is not a family friendly restaurant. It has a high customer rating of 5 out of 5. You can find it near the Crowne Plaza Hotel.*

***S2**: JJ's Pub is not family friendly, but it has a high customer rating of 5 out of 5. It is a restaurant near the Crowne Plaza Hotel.*

Why control discourse?

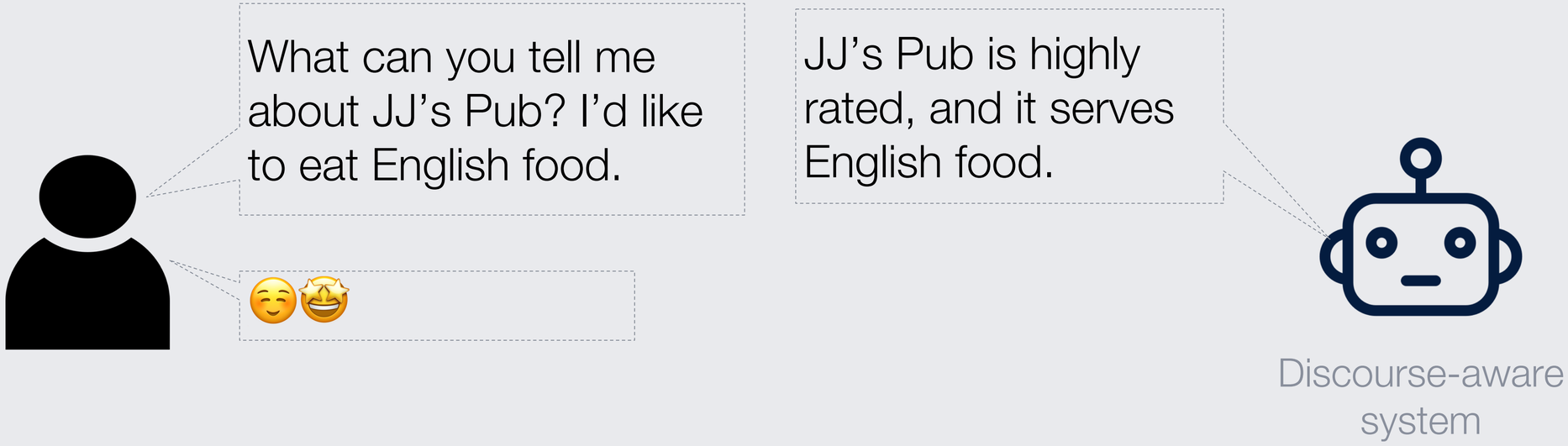
Increased **perceived naturalness** when user models influence discourse



(Lemon et al. 2004; Carenin and Moore, 2006; Walker et al. 2007; White et al. 2010; Demberg et al. 2011)

Why control discourse?

Prior work has shown increases in **perceived naturalness** when user models influence discourse



(Lemon et al. 2004; Carenin and Moore, 2006; Walker et al. 2007; White et al. 2010; Demberg et al. 2011)

Reed et al. (2018)

- Add input tokens to indicate target discourse structure
  - contrast
  - # of sentences

- Greatly improved accuracy of contrast expression and # sentences

- **But**: no way to control **which slots** gets contrasted!

How can we control the expression of discourse relations?

Compositional Meaning Representations

Proposed Representation Components

- **Arguments** e.g. `date_time, family_friendly`
  - Entities or slots to be mentioned
  - Can be nested (e.g. `date_time` contains `week_day`)

- **Dialog acts** e.g. INFORM, RECOMMEND
  - Contain arguments

- **Discourse relations** e.g. JUSTIFY, CONTRAST
  - Relationship between dialog acts / discourse relations

(Mann and Thompson, 1988; Rambow et al. 2001; Reiter and Dale, 2000; Walker et al. 2007)

Dialog Acts

DG:INFORM	Response with requested information
DG:ACK	Acknowledgement of a request
DG:CONFIRM	Confirmation before performing an action
DG:PROMPT	Asks the user if they want to perform a task, usually as a follow-up
DG:ERROR	When the assistant can't perform an action, usually due to an invalid argument
DG:REQUEST	When more information is needed to perform an action
DG:DISAMB	When arguments need to be disambiguated
DG:SELECT	When the user needs to choose from a list

Discourse Acts

DS:JOIN	Combines dialog acts, usually with a conjunction like “and”
DS:CONTRAST	Combines dialog acts that differ in some way, usually with a conjunction like “but”
DS:JUSTIFY	Used when one dialog act provides justification for another dialog act

Putting it all together

- Tree-structured representation
- Allows arbitrary nesting of relations and acts

```
[JOIN
 [CONTRAST
  [INFORM [name JJs’ Pub] [rating 5 out of 5] ]
  [INFORM [familyFriendly no ] ]
 ]
 [INFORM [eatType restaurant] [near Crowne Plaza Hotel] ]
]
```

Flat MR vs. Tree-structured MR

Flat MR	condition1[sunny] date_time1[this weekend] avg_high1[60s] low2[43] date_time2[Sunday evening] chance3[likely] wind_summary3[strong] date_time3[Saturday morning]
Our MR	INFORM [condition[sunny], date_time_range[colloquial[this weekend]]] CONTRAST [INFORM [avg_high[60s] date_time[[colloquial this weekend]]] INFORM [low[43] date_time[week_day[Sunday] colloquial[evening]]]] INFORM [chance[likely], wind_summary[heavy], date_time[week_day[Saturday] colloquial[morning]]]
Annotated Reference	[INFORM It'll be [condition sunny] throughout [date_time_range colloquial[this weekend]]. [CONTRAST [INFORM The high will be in the [avg_high 60s]]] , [INFORM but expect temperatures to drop as low as [avg_low 43 degrees] by [date_time [week_day Sund [colloquial evening]]]. [INFORM There's also [chance a chance of] [wind_summary strong winds] on [date_time [week_day Saturday] [colloquial morning]] .]

Data

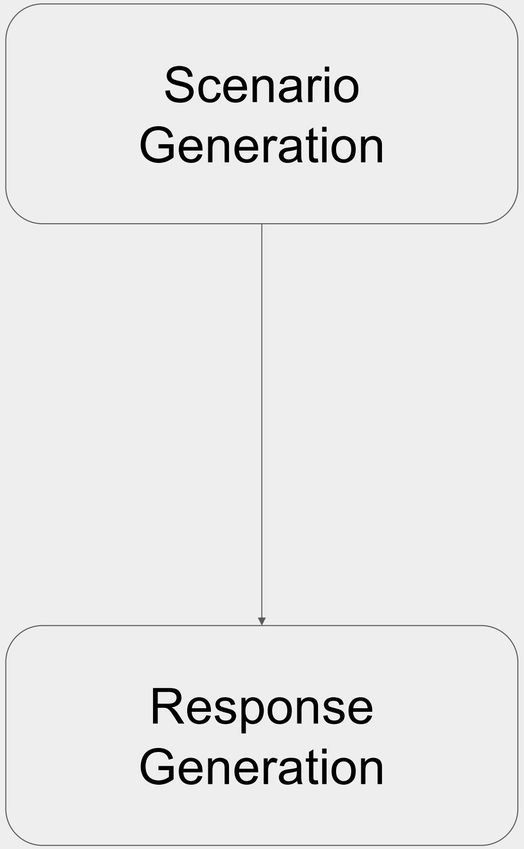
Dataset Creation

Rule-based

User query crowdsourcing → Compositional MR generation → Response generation and annotation → Response quality evaluation → **33K examples** for **weather** domain

https://github.com/facebookresearch/TreeNLG

Annotation Process



Q: Is it going to rain in Miami tomorrow?

S: [DS_CONTRAST
[DG_NO]
[DG_INFORM
[date_time tomorrow]
[location Miami]
[condition sunny]
]
]

[DS_CONTRAST [DG_NO No,] [DG_INFORM
it'll be [condition sunny] [date_time tomorrow]
in [location Miami]]] .

Modified E2E Dataset

- Used the Berkeley neural parser to automatically infer CONTRAST and JOIN relations in the E2E challenge dataset

- Used Slug2Slug token tagger and HarvardNLP latent segmentation model to annotate responses

Approach

Model Overview

Standard Seq2Seq model with attention

Input: Linearized tree structure

```
[CONTRAST [INFORM [condition rain ] [date_time [week_day Saturday ] ]
] [INFORM [cloud_coverage sunny ] [date_time [week_day Sunday ] ] ] ]
```

Output: Response with tree structure preserved

```
[CONTRAST [INFORM There'll be [condition rain ] on [date_time
[week_day Saturday ] ] ], but [INFORM [date_time [week_day Sunday ] ]
will be [cloud_coverage sunny ] ] ].
```

Semantic Correspondence

Leverage correspondence between input and output structures

```
MR: [CONTRAST [INFORM [condition rain ] [date_time [week_day Saturday
] ] ] [INFORM [cloud_coverage sunny ] [date_time [week_day Sunday ] ]
] ]

S: [JOIN [INFORM There'll be [condition rain ] on [date_time
[week_day Saturday ] ] ], and [INFORM [date_time [week_day Sunday ] ]
will be [cloud_coverage sunny ] ] ].
```

Semantic Correspondence

Leverage correspondence between input and output structures

```
[CONTRAST [INFORM [condition rain ] [date_time [week_day Saturday ] ]
] [INFORM [cloud_coverage sunny ] [date_time [week_day Sunday ] ] ] ]
```

Tree accuracy metric for semantic correctness

```
[JOIN [INFORM There'll be [condition rain ] on [date_time [week_day
Saturday ] ] ], and [INFORM [date_time [week_day Sunday ] ] will be
[cloud_coverage sunny ] ] ].
```

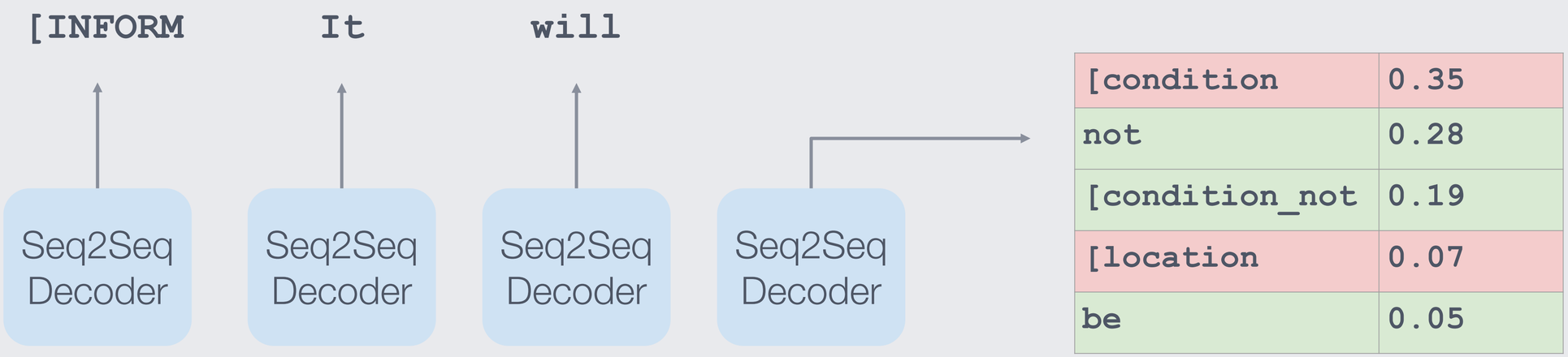
Key Model Consideration

- Lexicalization vs. delexicalization
  - i.e., “It will be [condition sunny]” vs. “It will be [condition _condition_]”
  - Free surface form realization vs vocabulary sparsity
- Coarse-grained vs. fine-grained
  - Coarse: [INFORM [ARG_CONDITION sunny] [DATE today]].
  - Fine: [INFORM [ARG_CONDITION_SUNNY sunny] [DATE today]].
  - Fine-grained arguments provides us better correctness control, but also increase the vocab space.

Constrained Decoding

Idea: During beam search, invalidate tokens that would result in responses that don't match the input structure (based on the nonterminals in the output)

MR: [**INFORM** [**condition_not** rain] [**date_time** [**colloquial** today]]]



Constrained Decoding (contd.)

Mask tokens that invalidate constraints:

- validity of child node
- coverage of all children before closing a node
- allow *aggregation* of similar attributes

```
MR: [INFORM [location Menlo Park ] [condition sunny] ]
[INFORM [location Menlo Park ] [temp 72 ] ]

S: [INFORM [location Menlo Park ] will be [condition sunny
] ] [INFORM with a temperature of [temp 72 ] ].
```

Tree2Seq

Why Tree2Seq?

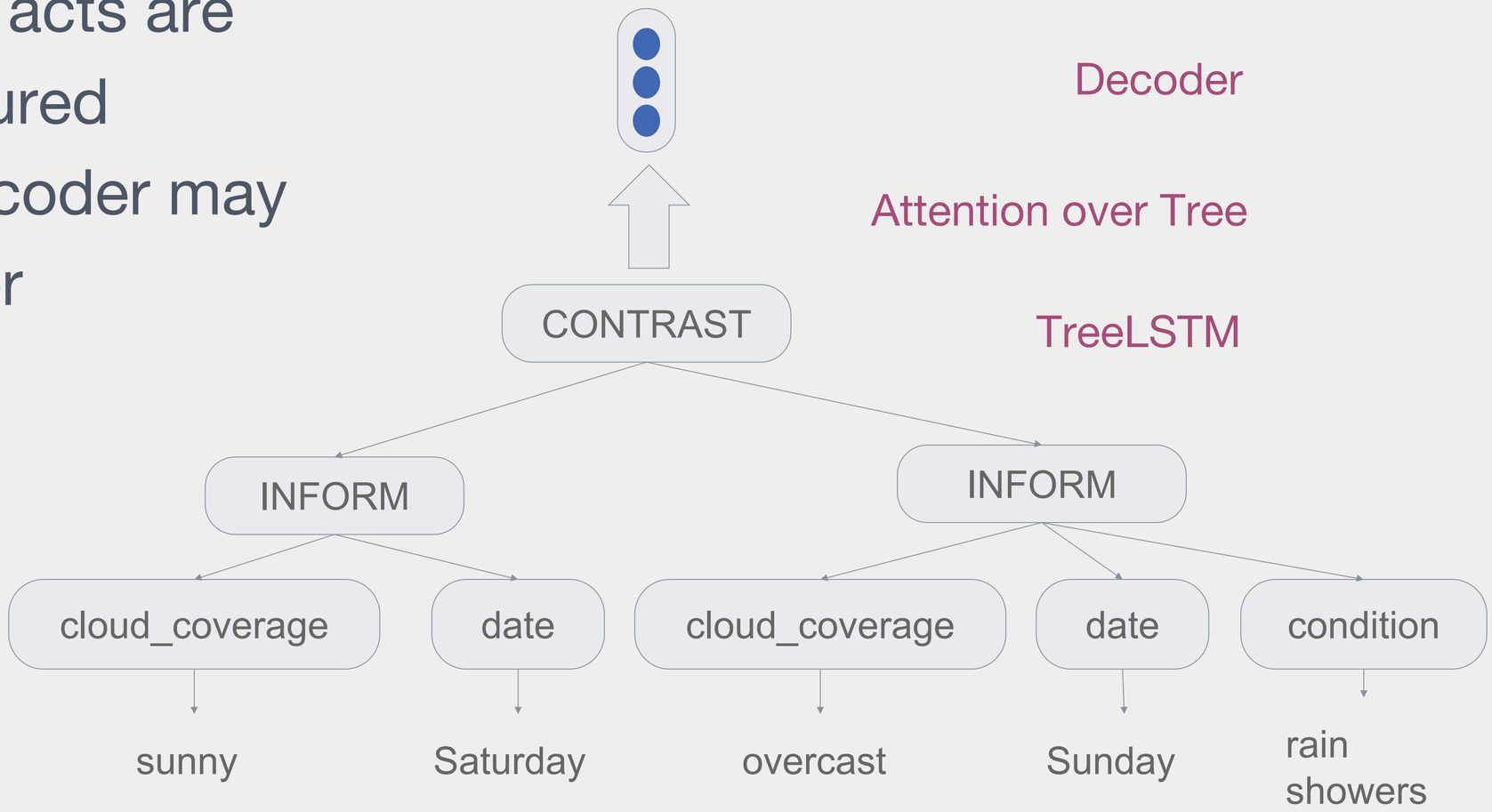
Would combining a better learning and decoding algorithm leveraging tree structures lead to better effectiveness?

Tree-structured MR:

Our MR	INFORM [condition[sunny], date_time_range[colloquial[this weekend]]] CONTRAST [INFORM [avg_high[60s] date_time[[colloquial this weekend]]] INFORM [low[43] date_time[week_day[Sunday] colloquial[evening]]]] INFORM [chance[likely], wind_summary[heavy], date_time[week_day[Saturday] colloquial[morning]]]

Tree-to-sequence model

- Dialog and discourse acts are inherently tree-structured
- A structure-aware encoder may learn semantics better



Tree-to-sequence Model

- N-ary Tree-based encoder (Tai et al. ACL'15)

$$\mathbf{h}_k^p = f_{\text{tree}}(\{\mathbf{h}_k^{c_1}, \ldots, \mathbf{h}_k^{c_N}\})$$

- Tree inputs are hard to parallelize
  - Developed an iterative bottom-to-up traversal algorithm to support batch forward/backward
  - Achieved 5-10x speedup in training and inference

Tree-to-sequence Model

- Structured-aware Decoder
  - Initialize decoder state as root hidden state

$$\mathbf{s}_1 = \mathbf{h}_{\text{root}}$$

- Incorporate tree structures into decoding

$$\widehat{\mathbf{s}_j} = \tanh(\mathbf{W}_d \cdot [\mathbf{s}_j; \mathbf{d}_j; \widehat{\mathbf{s}_{j-1}}] + \mathbf{b}_d)$$

$$P(y_i|\mathbf{y}_{<i}, x) = (\mathbf{W}_s \cdot \widehat{\mathbf{s}_j} + \mathbf{b}_s)$$

Experiments and Results

Models

- **S2S-FLAT**
  - Input and output contain a flat list of arguments
- **S2S-TOKEN**
  - Inspired by Reed et al. 2018, input is similar to Flat but has tokens indicating # of Contrast and Join relations
- **S2S-TREE**
  - Input and output are linearized tree representations
- **S2S-CONSTR**
  - TREE with constrained decoding
- **T2S**
  - Tree2seq model with linearized tree representations
- **T2S-CONSTR**
  - Tree2seq + constrained decoding

Metrics

- Automatic
  - BLEU
  - Tree accuracy
- Human evaluation
  - Grammaticality
  - Semantic correctness
  - Disc. correctness (semantic correctness measured on challenging subset)

Results

Model	E2E					Weather				
Metric	BLEU	TreeAcc	Gram	Corr	Disc	BLEU	TreeAcc	Gram	Corr	Disc
S2S-FLAT	0.6360	-	94.03	63.85	30.87	0.7455	-	98.77	77.09	79.04
S2S-TOKEN	0.7441‡	-	92.29	69.02†	42.29†	0.7493*	-	96.7	81.56†	83.93†
S2S-TREE	0.7458‡	94.86	93.59	83.85†	54.35†	0.7612*	92.5	95.26	87.61†	85.97†
S2S-CONSTR	**0.7469‡**	**99.25**	**94.33**	**85.89†**	**66.09†**	**0.7660***	**96.92**	95.30	**91.82†**	**93.44†**

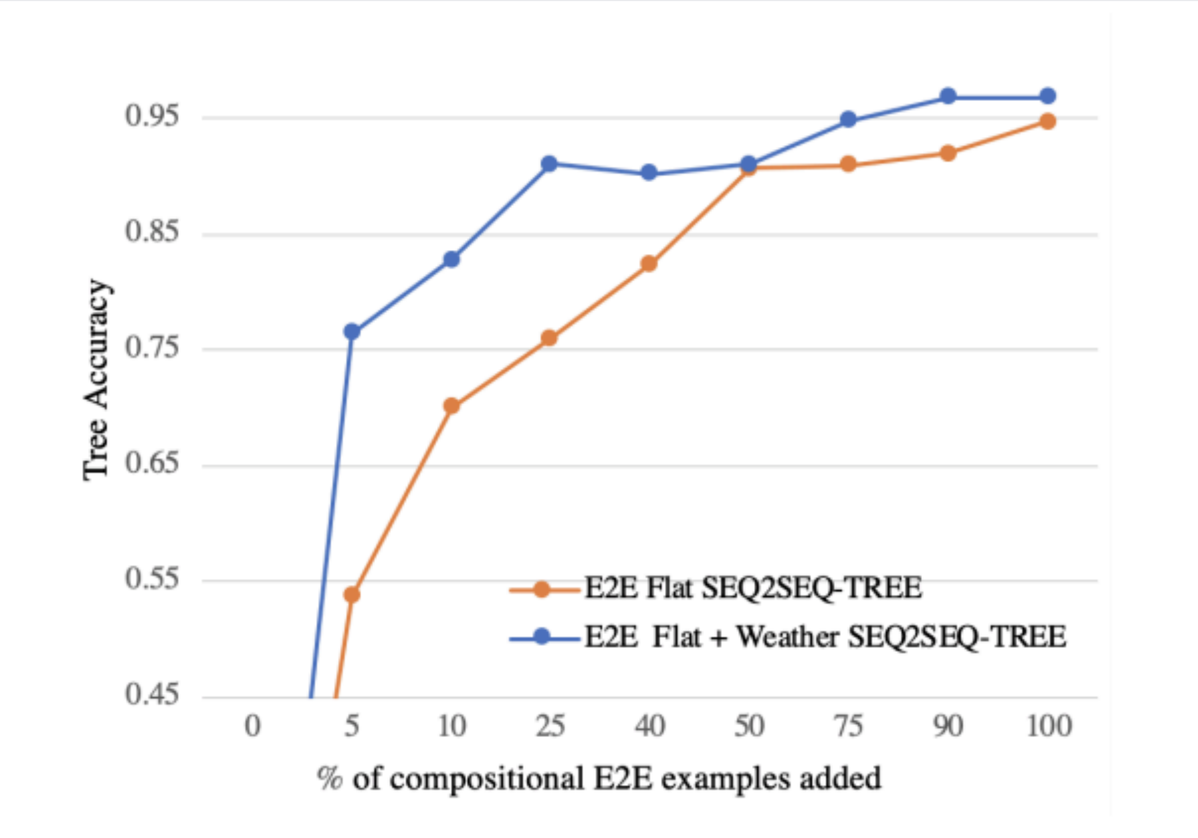
Tree-based representations result in responses that are more consistent with desired discourse structure

Constrained decoding improves semantic correctness further

Observations

- Constrained decoding can still result in incorrect generations
  - Model generates surface forms without generating the non-terminal
  - [restaurant <name>] is not family friendly and serves [cuisine Indian] food.
- Occasional (<1%) stuttering due to constrained decoding
- Unnatural phrasings
  - *Yes, you should wear [ATTIRE an umbrella]*

Generalization Study



Trained two models:

1) *On E2E flat data*
2) *On E2E flat data + all Weather data*

Gradually add more compositional E2E examples to training

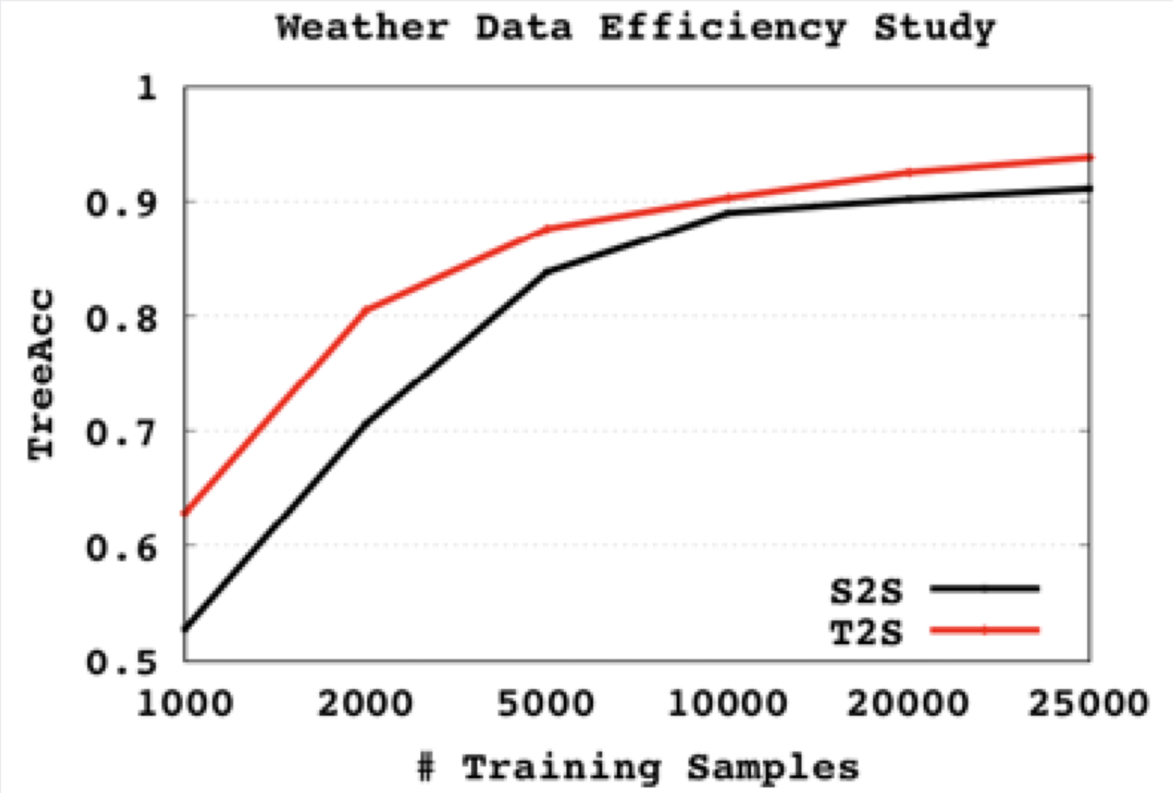
Tree2Seq Results

Model	E2E					Weather				
Metric	BLEU	TreeAcc		Gram.	Corr.	BLEU	TreeAcc		Gram.	Corr.
	-	NoDisc	Disc	-	-	-	NoDisc	Disc	-	-
S2S	74.58	99.68	95.28	93.59	83.85	76.75	96.62	83.30	94.17	87.40
S2S-Constr	74.69	**99.89**	97.78	94.33	**85.89**	77.45	98.52	91.61	94.20	90.40
Our Approaches										
T2S	**74.75**	**99.89**	96.96	94.83	84.66	**77.86**	97.1	88.80	**94.55**	89.75
T2S-Constr	74.63	99.84	**98.60**	**94.68**	85.68	77.82	**99.11**	**94.13**	94.14	**91.84**

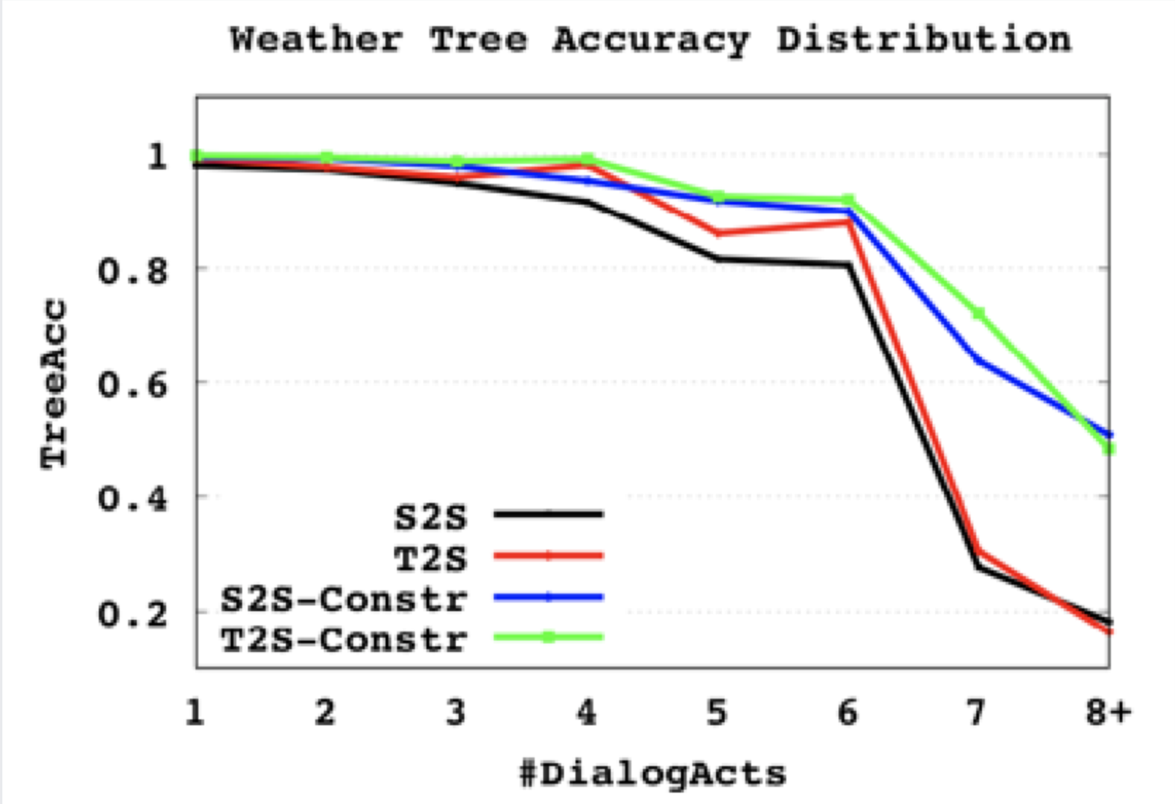
- No big differences in BLEU and Grammaticality score
- Tree2Seq + constrained decoding achieved the best tree accuracy
- Human correctness highly correlates to tree accuracy

Model Effectiveness

Tree Acc by % data



Tree Acc by # DialogActs



Model Samples

Column	G/T/C	Text
MR Input	-	Join [INFORM [area[riverside] name[_name_] price_range[high] eat_type[restaurant] [near _near_]] CONTRAST [INFORM [rating[1 out of 5] family_friendly_no[no]] INFORM [food_italian[_food_italian_]]]]
Annotated Reference	-	[Join [INFORM There is a [price_range high] price range [eat_type restaurant] [name name] in [area riverside] near [_near_]] [CONTRAST [INFORM It is [family_friendly not kids friendly] and [rating has low] customer ratings] but [INFORM they serve [food_italian Italian food] .]]]
S2S	1/0/0	[Join [INFORM [name _name_] is a [price_range high] priced [eat_type restaurant] in the [area riverside] area near [near _near_] .] [CONTRAST [INFORM is [family_friendly_no not children friendly] ,] but [INFORM it has a customer rating of [rating 1 out of 5] .]]]
T2S	1/0/0	[Join [INFORM [name _name_] is a [price_range high] priced [eat_type restaurant] in the [area riverside] area near [near _near_] .] [CONTRAST [INFORM it has a customer rating of [rating 1 out of 5] ,] but [INFORM is [family_friendly_no not children friendly] and serves [food_italian Italian] food.]]]
T2S-Constr	1/1/1	[Join [INFORM [name _name_] is a [price_range high] priced [eat_type restaurant] in the [area riverside] area near [near _near_] .] [CONTRAST [INFORM it is [family_friendly not children friendly] and has a [rating 1 out of 5] rating] but [INFORM it serves [food_italian Italian] food.]]]

- G/T/C stands for grammatical/tree accuracy/correct (human annotation).

Grammaticality Filter

Correctness filter motivation

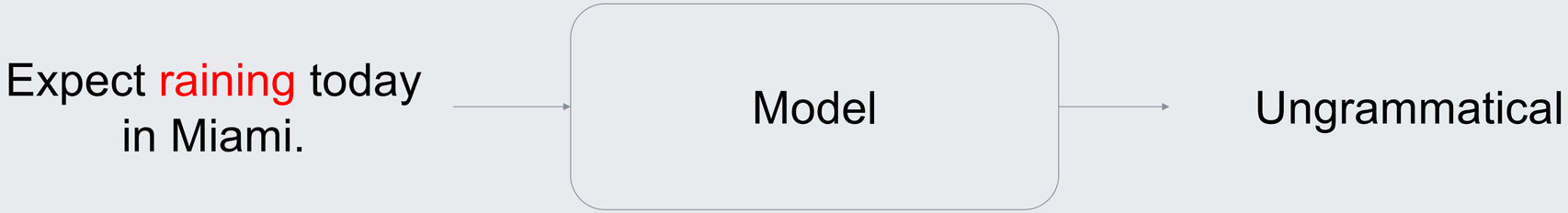
Input	**[INFORM [CONDITION sunny] [DATE_TIME today]]**
Correct Response	It will be sunny today.
Slot Drop	It will be sunny.
Hallucination	It will be cloudy today.
Added information	It will be sunny and 45 F today.

Correctness filter

Idea: Use tree structure predicted by model to automatically detect incorrect responses

Input	**[INFORM [CONDITION sunny] [DATE_TIME today]]**
Correct Response with non-terminals	[INFORM It will be [CONDITION sunny] [DATE_TIME today]] .
Correct Response with fine-grained non-terminals	[INFORM It will be [CONDITION_SUNNY sunny] [DATE_TIME today]] .
Response with Hallucination and fine-grained non-terminals	[INFORM It will be [CONDITION_CLOUDY cloudy] [DATE_TIME today]].

Grammaticality filter



- DocNN: Biased for precision, 0.85 recall at 0.98 precision
- BERT: 0.99 precision at 0.90 recall
- Future work:
  - Develop a domain-agnostic grammaticality filter
  - Distill BERT to smaller model for product serving.

Conclusions

- Shown that tree structured representations can greatly improve controllability of generated text
- Introduced a simple constrained decoding technique that is only applied to the decoder
- Introduced a tree-to-sequence model for better structure-aware learning
- Released a conversational NLG corpus for the weather domain
- Introduced grammatical filtering approach

Check out our dataset and code:

https://github.com/facebookresearch/TreeNLG

Future Work

- Add constraints into training (trainable beam search)
- Add LM/reranking approach for better grammaticality and correctness
- Adversarial approaches to improve model robustness
- Add multi-modal contexts into generation
- Personalized NLG model
- Cross-domain and cross-lingual generalization
- Data Augmentation to leverage public data for domain bootstrapping

Thank you!

https://github.com/facebookresearch/TreeNLG

facebook