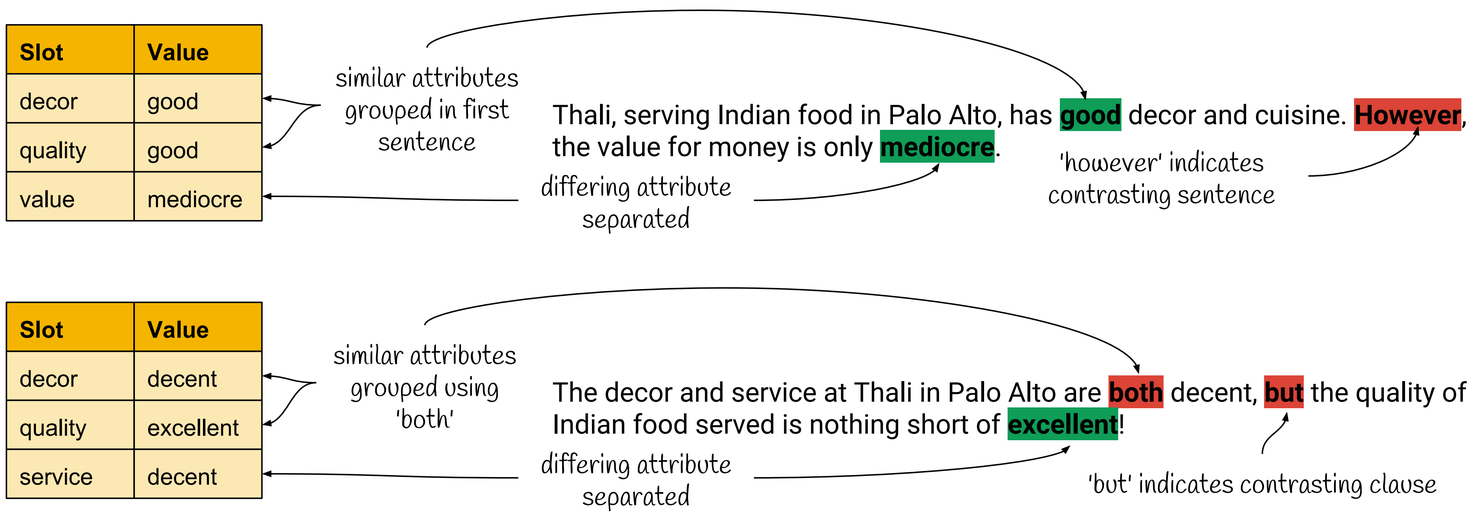
To Plan or not to Plan?

Discourse planning in slot-value informed sequence-to-sequence models for language generation

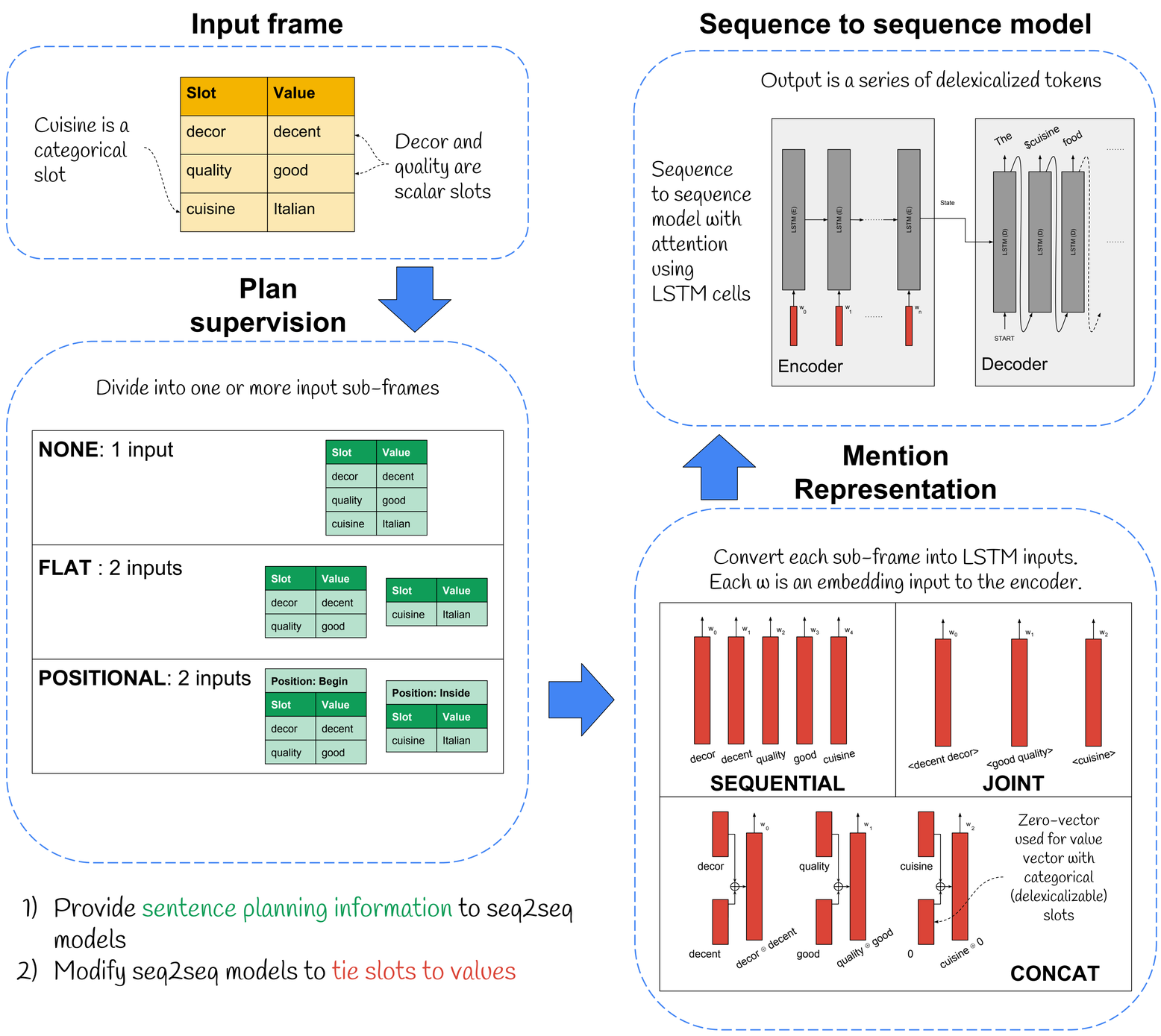
Neha Nayak, Dilek Hakkani-Tür, Marilyn Walker, Larry Heck
nayakn@google.com, dilek@ieee.org, mawalker@ucsc.edu, larry.heck@ieee.org

Planning in Neural Models



- Slot values can affect (1) **grouping of mentions** into sentences (2) **tokens** surrounding mentions
- We apply concepts of **sentence planning** to neural models, specifically to slot aggregation
- We condition generation on **slot values**, handling scalar values which **cannot be delexicalized**

Mention Representation and Plan Supervision



1) Provide sentence planning information to seq2seq models
2) Modify seq2seq models to tie slots to values

Data

Examples elicited for all possible combinations of scalar slots, with all possible value assignments. Total: 1662 examples.

	Slot	Possible value
categorical	Name	Au Midi
	Neigborhood	Midtown
	Cuisine	French
scalar	Decor	excellent
	Food quality	good
	Service	decent
	Value for money	mediocre

Evaluation

- Objective: Slot precision, slot recall, scalar precision (% slots with correct values)
- Subjective: Naturalness, Syntax, Overall ratings (comparing 3 utterances)
- Baselines: Utterance sampled from training set; HUMAN-S: with correct slots and values; HUMAN-G: with correct slots and any values

Results

Results for mention representation and plan supervision. * indicates CONCAT values significantly better than SEQ; † indicates significantly better than HUMAN-S (both $p < 0.05$). Bold values of POSITIONAL are significantly better than NONE. ($p < 0.05$).

		Slot Prec.	Scalar Prec.	Slot Rec.	Naturalness	Syntax	Overall	# uniq. sents
Mention representation	SEQ	1.0	0.6	92.24%	2.72	3.04	2.75	100
	JOINT	1.0	0.86	86.26%	2.42	2.75	2.41	101
	CONCAT	1.0	0.98	**95.33%***	**2.80**†	2.98	**2.87***	126
	HUMAN-G	1.0	1.0	94.98%	2.65	2.93	2.76	313
	HUMAN-S	1.0	1.0	95.29%	2.64	2.87	2.79	325
Plan supervision	NONE	-	-	97.22%	2.69	2.80	2.68	126
	FLAT	-	-	96.99%	2.68	2.93	2.74	152
	POSITIONAL	-	-	96.99%	2.74	**3.02**	**2.81**	144
	HUMAN-S	-	-	96.01%	2.72	2.91	2.83	325

Observations

- Model gravitates towards 'safe', commonly occurring, grammatical occurrences
- Scores higher than human baseline on average for naturalness, syntax
- Human baselines still produce higher diversity: CONCAT and HUMAN-S produced a total of 126 and 325 unique sentences in the test set, respectively.
- 4 most frequent sentence plans account for 524 utterances (of total 1662 utterances)

Conclusions

- Conditioning on slot values helps tackle scalar valued slots
- Adding planning information can help increase diversity and perceived quality