Findings of the E2E NLG Challenge

Ondřej Dušek, **Jekaterina Novikova** and **Verena Rieser**

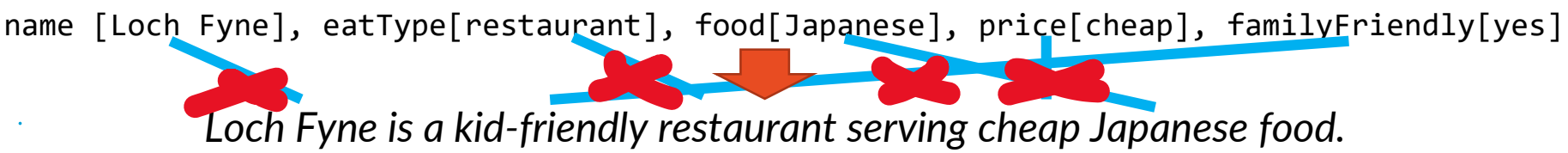
Interaction Lab, Heriot-Watt University

INLG, Tilburg

7 November 2018

E2E NLG Challenge

- Task: generating restaurant recommendations
  - simple input MR, no content selection (as in dialogue systems)
- **New neural NLG:** promising, but so far limited to small datasets
- **"E2E" NLG**: Learning from just pairs of MRs + reference texts
  - **no alignment** needed → easier to collect data



- **Aim:** Can new approaches do better if given more data?

E2E Dataset

Novikova et al.
SIGDIAL 2017
[ACL W17-5525]

HERIOT WATT UNIVERSITY

- Well-known restaurant domain
- **Bigger** than previous sets
  - **50k MR+ref pairs** (unaligned)

	Instances	MRs	Refs/MR	Slots/MR	W/Ref	Sent/Ref
E2E	**51,426**	**6,039**	**8.21**	**5.73**	**20.34**	**1.56**
SF Restaurants	5,192	1,914	1.91	2.63	8.51	1.05
Bagel	404	202	2.00	5.48	11.55	1.03

- **More diverse** & natural
  - partially collected using pictorial MRs
  - **noisier**, but compensated by more refs per MR

```
name [Loch Fyne], eatType[restaurant],
food[Japanese], price[cheap],kid-friendly[yes]
```

Loch Fyne is a kid-friendly restaurant serving cheap Japanese food.

Serving low cost Japanese style cuisine, Loch Fyne caters for everyone, including families with small children.

E2E Challenge timeline

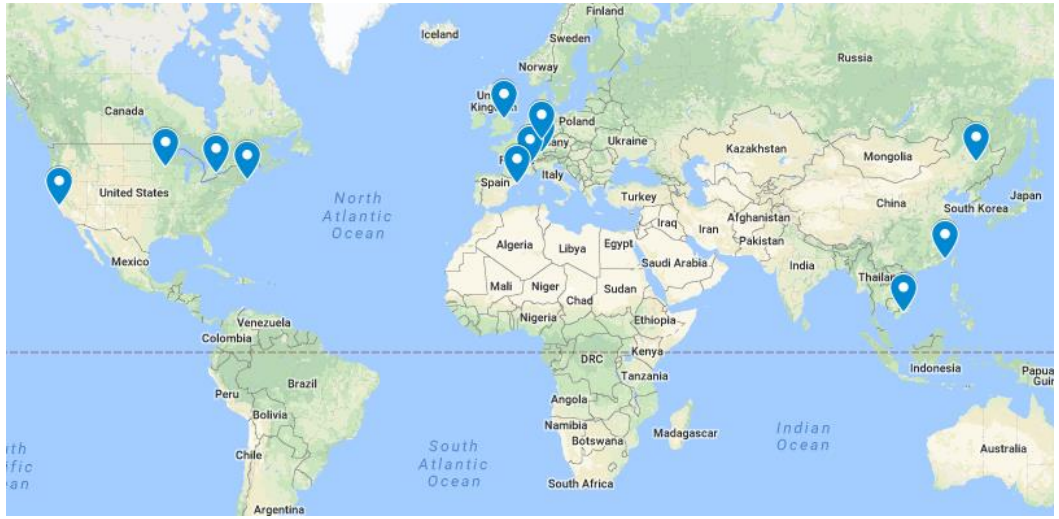
HERIOT WATT UNIVERSITY

- Mar ’17: Training data released
- Jun ’17: Baseline released
- Oct ’17: Test MRs released (16th), submission deadline (31st)
- Dec ’17: Evaluation results released
  Technical papers submission
- Mar ’18: Final technical papers + full data released
- Nov ’18: Results presented, outputs & ratings released

http://bit.ly/e2e-nlg

E2E Participants

- 17 participants (⅓ from industry), 62 submitted systems
  - success!



- 3 withdrew after automatic evaluation
  - → 14 participants
  - 20 primary systems + baseline for human evaluation

Participants: Architectures

- Seq2seq: 12 systems + baseline
  - many variations & additions
- Other fully data-driven: 3 systems
  - 2x RNN with fixed encoder
  - 1x linear classifiers pipeline
- Rule/grammar-based: 2 systems
  - 1x rules, 1x grammar
- Templates: 3 systems
  - 2x mined from data, 1x handcrafted

TGEN	HWU (baseline)	*seq2seq + reranking*
SLUG	UCSC Slug2Slug	*ensemble seq2seq + reranking*
SLUG-ALT	UCSC Slug2Slug	*SLUG + data selection*
TNT1	UCSC TNT-NLG	*TGEN + data augmentation*
TNT2	UCSC TNT-NLG	*TGEN + data augmentation*
ADAPT	AdaptCentre	*preprocessing step + seq2seq + copy*
CHEN	Harbin Tech (1)	*seq2seq + copy mechanism*
GONG	Harbin Tech (2)	*TGEN + reinforcement learning*
HARV	HarvardNLP	*seq2seq + copy, diverse ensembling*
ZHANG	Xiamen Uni	*subword seq2seq*
NLE	Naver Labs Eur	*char-based seq2seq + reranking*
SHEFF2	Sheffield NLP	*seq2seq*
TR1	Thomson Reuters	*seq2seq*
SHEFF1	Sheffield NLP	*linear classifiers trained with LOLS*
ZHAW1	Zurich Applied Sci	*SC-LSTM RNN LM + 1st word control*
ZHAW2	Zurich Applied Sci	*ZHAW1 + reranking*
DANGNT	Ho Chi Minh Ct IT	*rule-based 2-step*
FORGE1	Pompeu Fabra	*grammar-based*
FORGE3	Pompeu Fabra	*templates mined from data*
TR2	Thomson Reuters	*templates mined from data*
TUDA	Darmstadt Tech	*handcrafted templates*

E2E Generation Challenges

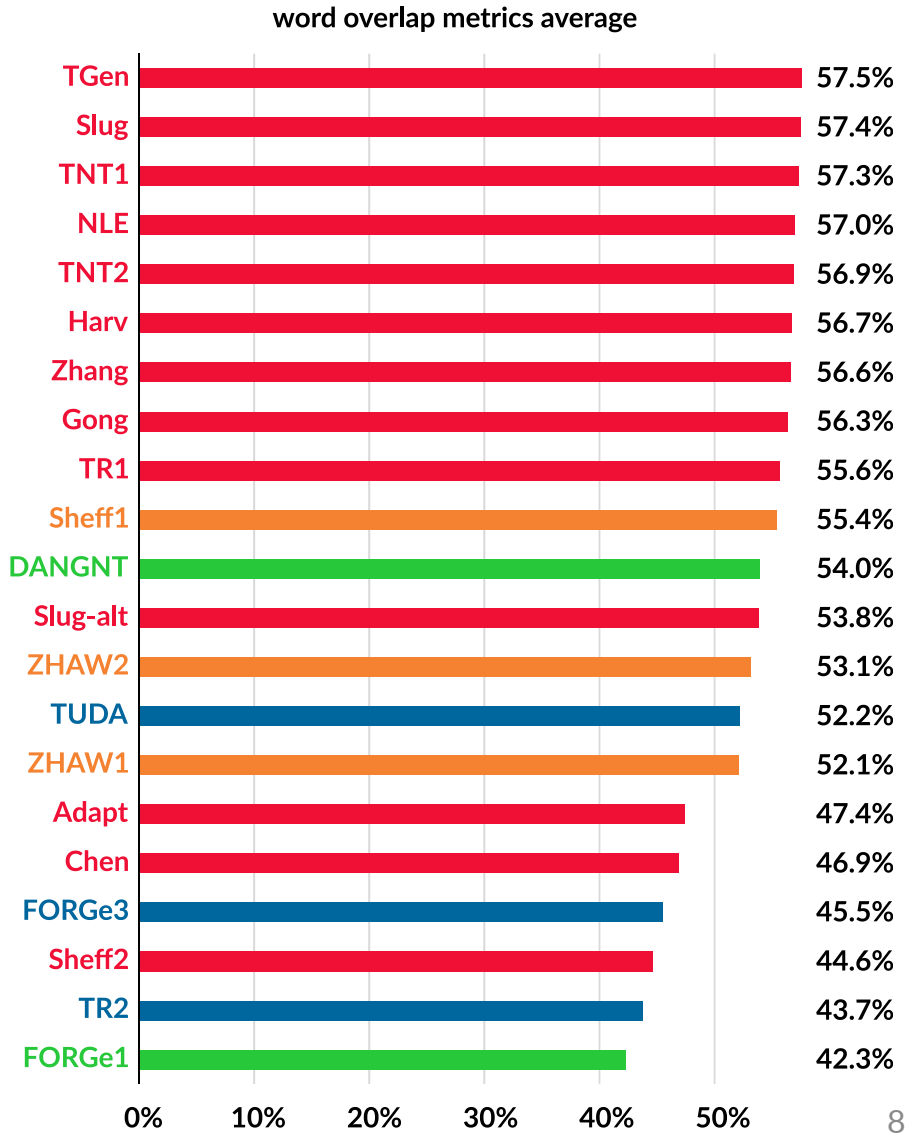
HERIOT WATT UNIVERSITY

- **Open vocabulary** (restaurant names)
  - delexicalization – placeholders
  - seq2seq: copy mechanisms, subword/character level
- **Semantic control** (realizing all attributes)
  - template/rule-based, SHEFF1: given by architecture
  - seq2seq: beam reranking – MR classification/alignment (some systems)
- **Output diversity**
  - data augmentation / data selection
  - diverse ensembling (HARV)
  - preprocessing steps (ZHAW1, ZHAW2)

Automatic evaluation: Word-overlap metrics

- Several commonly used
  - BLEU, NIST, METEOR, ROUGE, CIDEr
- Scripts provided
  - **http://bit.ly/e2e-nlg**
- Baseline very strong
- Seq2seq systems best, but some bad

- Segment-level correlation vs. humans weak (<0.2)



Human evaluation

Novikova et al.
NAACL 2018
[ACL N18-2012]

HERIOT WATT UNIVERSITY

- Criteria: **naturalness** + **overall quality**
  - separate collection to lower correlation
  - input MR not shown to workers evaluating naturalness
- **RankME** – relative comparisons & continuous scales
  - we found it to increase consistency vs. Likert scales / single ratings
- **TrueSkill** (Sakaguchi et al. 2014)– fewer direct comparisons needed
  - significance clusters established by bootstrap resampling

Human evaluation – example (Quality)

System	Output	Rank	Score
	name[Cotto], eatType[coffee shop], near[The Bakers]		
TR2	*Cotto is a coffee shop located near The Bakers.*	1	100
SLUG-ALT	*Cotto is a coffee shop and is located near The Bakers*	2	97
TGEN	*Cotto is a coffee shop with a low price range. It is located near The Bakers.*	3-4	85
GONG	*Cotto is a place near The Bakers.*	3-4	85
SHEFF2	*Cotto is a pub near The Bakers.*	5	82
	name[Clowns], eatType[coffee shop], customer rating[3 out of 5], near[All Bar One]		
SHEFF1	*Clowns is a coffee shop near All Bar One with a customer rating of 3 out of 5.*	1-2	100
ZHANG	*Clowns is a coffee shop near All Bar One with a customer rating of 3 out of 5 .*	1-2	100
FORGE3	*Clowns is a coffee shop near All Bar One with a rating 3 out of 5.*	3	70
ZHAW2	*A coffee shop near All Bar One is Clowns. It has a customer rating of 3 out of 5.*	4	50
SHEFF2	*Clowns is a pub near All Bar One.*	5	20

Human evaluation results

- 5 clusters each, clear winner
- Naturalness:
  Seq2seq dominates
  - diversity-attempting systems penalized
- Quality: more mixed
  - 2nd cluster – all archs.
  - bottom clusters: seq2seq w/o reranking
- Overall winner: **SLUG**

Naturalness

#	Rank	System
1	1-1	SHEFF2
2	2-3	SLUG
	2-4	CHEN
	3-6	HARV
	4-8	NLE
	4-8	TGEN
	5-8	DANGNT
	5-10	TUDA
	7-11	TNT2
	9-12	GONG
	9-12	TNT1
	10-12	ZHANG
3	13-16	TR1
	13-17	SLUG-ALT
	13-17	SHEFF1
	13-17	ZHAW2
	15-17	ZHAW1
4	18-19	FORGE1
	18-19	ADAPT
5	20-21	TR2
	20-21	FORGE3

Quality

#	Rank	System
1	1-1	SLUG
2	2-4	TUDA
	2-5	GONG
	3-5	DANGNT
	3-6	TGEN
	5-7	SLUG-ALT
	6-8	ZHAW2
	7-10	TNT1
	8-10	TNT2
	8-12	NLE
	10-13	ZHAW1
	10-14	FORGE1
	11-14	SHEFF1
	11-14	HARV
3	15-16	TR2
	15-16	FORGE3
4	17-19	ADAPT
	17-19	TR1
	17-19	ZHANG
5	20-21	CHEN
	20-21	SHEFF2

E2E: Lessons learnt

HERIOT WATT UNIVERSITY

- (not strictly controlled setting!)
- **Semantic control** (realize all slots)– crucial for seq2seq systems
  - beam reranking works well, attention-only performs poorly
- **Open vocabulary** – delexicalization easy & good
  - other (copy mechanisms, sub-word/character models) also viable
- **Diversity** – hand-engineered systems seem better
  - options for seq2seq: diverse ensembling, sampling…
  - might hurt naturalness
- **Best method**: rule-based or seq2seq with reranking

Thanks

HERIOT WATT UNIVERSITY

- Get E2E NLG data & metrics & system outputs with rankings:
  http://bit.ly/e2e-nlg
- Contact us:

o.dusek@hw.ac.uk
@tuetschek

v.t.rieser@hw.ac.uk
@verena_rieser

- More detailed results analysis coming soon (on arXiv)!

E2E dataset: Novikova et al. SIGDIAL '17 [ACL W17-5525]

RankME eval: Novikova et al. NAACL '18 [ACL N18-2012]

Automatic evaluation: Textual metrics

HERIOT WATT UNIVERSITY

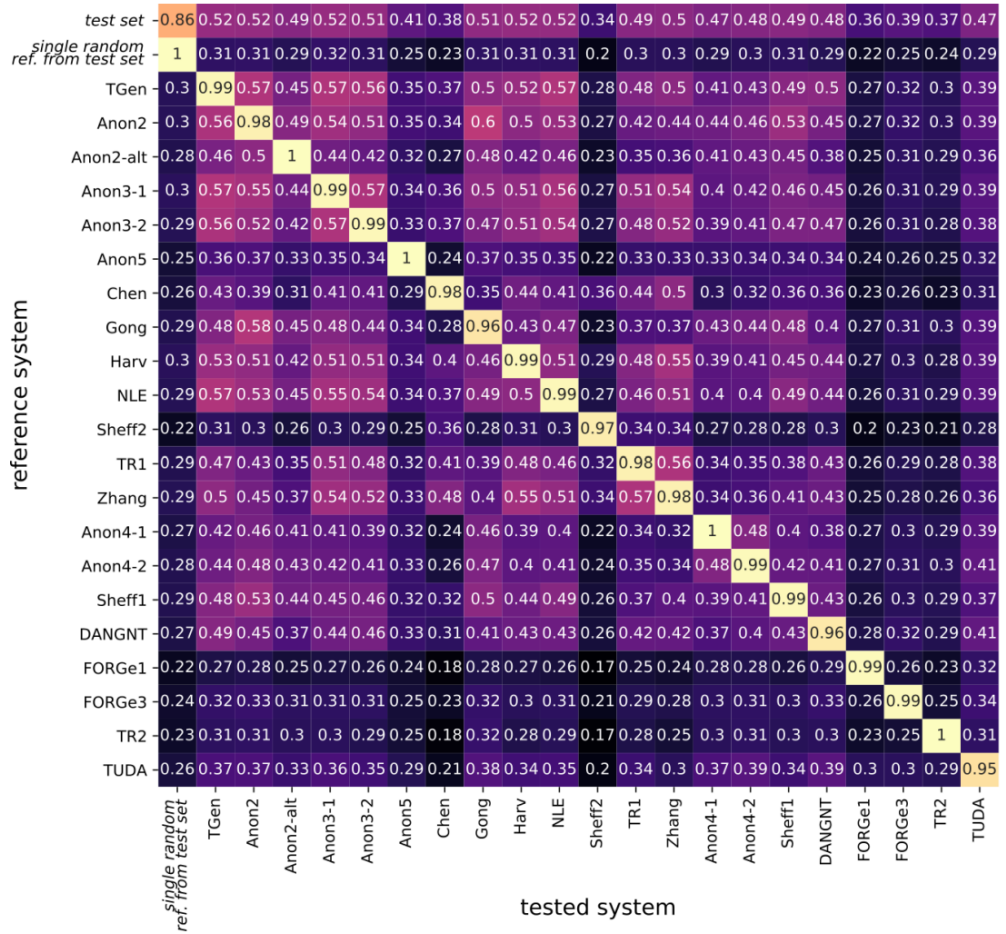
- Same diversity/complexity metrics used to evaluate the dataset

- Seq2seq-based systems – typically less syntactic complexity
- Rare words ratio typically same as in data (except FORGE1)
- Highest MSTTR:
  - rule/grammar-based systems
  - systems aiming at diversity (ZHAW1, ZHAW2, ADAPT, SLUG-ALT)
- Data-driven systems: shorter outputs than rule-based
  - low-performing seq2seq: very short outputs (CHEN, SHEFF2)

% D-Level0-2		% D-Level6-7		Rare words (LS2)		MSTTR-50		Average length	
ZHANG	88.98	SHEFF1	40.00	FORGE1	0.67	*train+dev set*	0.69	TUDA	30.05
TNT2	85.80	FORGE1	34.77	SHEFF2	0.61	TR2	0.63	FORGE1	26.73
TNT1	83.84	SLUG-ALT	24.97	ZHAW1	0.59	FORGE1	0.62	TR2	26.00
GONG	82.69	ZHAW1	23.84	CHEN	0.58	ADAPT	0.61	ZHAW1	25.05
SLUG	81.53	FORGE3	18.87	TR2	0.57	*test set*	0.58	ZHAW2	24.66
TR1	80.39	TR2	18.52	FORGE3	0.57	ZHAW1	0.56	DANGNT	23.67
DANGNT	79.66	ZHAW2	16.93	*test set*	0.57	ZHAW2	0.56	GONG	23.43
NLE	79.42	GONG	16.90	ADAPT	0.56	FORGE3	0.55	FORGE3	23.10
CHEN	78.99	*train+dev set*	15.44	ZHAW2	0.56	DANGNT	0.53	ADAPT	22.93
HARV	76.84	*test set*	14.64	HARV	0.56	SLUG-ALT	0.52	SLUG-ALT	22.89
SHEFF2	76.53	SLUG	11.30	TNT2	0.56	TUDA	0.52	TNT1	22.83
TGEN	76.42	TUDA	10.48	ZHANG	0.56	TGEN	0.50	*test set*	22.45
ADAPT	71.56	ADAPT	8.80	DANGNT	0.55	SLUG	0.50	TGEN	22.45
test set	67.80	TNT1	8.05	TGEN	0.54	HARV	0.49	SLUG	22.18
train+dev set	65.92	HARV	6.82	SHEFF1	0.54	SHEFF1	0.49	TNT2	21.89
FORGE3	65.36	TGEN	6.50	NLE	0.54	NLE	0.49	NLE	21.74
TR2	63.03	NLE	5.08	TR1	0.54	TNT1	0.49	HARV	21.47
TUDA	62.43	TR1	5.01	GONG	0.53	TNT2	0.49	SHEFF1	21.11
FORGE1	61.65	DANGNT	4.62	TUDA	0.52	TR1	0.47	TR1	20.93
ZHAW1	60.03	TNT2	4.13	TNT1	0.52	GONG	0.46	*train+dev set*	19.41
SLUG-ALT	59.06	SHEFF2	2.08	*train+dev set*	0.52	ZHANG	0.45	ZHANG	19.05
ZHAW2	56.52	ZHANG	1.95	SLUG-ALT	0.51	CHEN	0.42	SHEFF2	15.68
SHEFF1	37.93	CHEN	0.99	SLUG	0.51	SHEFF2	0.41	CHEN	14.67

Output similarity

- word-overlap metrics
  - systems against each other
- seq2seq most similar
  - except low-performing
- lower similarity for diversity-attempting
- lower similarity for template/rule-based



System	Mean
♡TGen	0.46
♡Slug	0.46
♡TNT1	0.46
♡NLE	0.45
♡TNT2	0.45
♡Gong	0.44
♡Harv	0.44
♢Sheff1	0.42
♡Zhang	0.42
♣DANGNT	0.42
♡TR1	0.42
♢ZHAW2	0.41
♡Slug-ALT	0.40
♢ZHAW1	0.39
♠TUDA	0.39
♡Adapt	0.34
♡Chen	0.34
♠FORGe3	0.32
♠TR2	0.31
random test set ref.	0.31
♣FORGe1	0.29
♡Sheff2	0.28

E2E Dataset: domain

HERIOT WATT UNIVERSITY

- Simple, well-known: restaurant information
- 8 attributes (slots)
  - most enumerable
  - 2 open: name/near (restaurant names)

Attribute	Data Type	Example value
name	verbatim string	*The Eagle, ...*
eatType	dictionary	*restaurant, pub, ...*
familyFriendly	boolean	*Yes / No*
priceRange	dictionary	*cheap, expensive, ...*
food	dictionary	*French, Italian, ...*
near	verbatim string	*market square, Cafe Adriatic, ...*
area	dictionary	*riverside, city center, ...*
customerRating	enumerable	*1 of 5 (low), 4 of 5 (high), ...*

- Aim: more varied, challenging texts than previous similar sets

E2E Data collection

Novikova et al. INLG 2016
[ACL W16-6644]

HERIOT WATT UNIVERSITY

- Crowdsourcing on CrowdFlower
- Combination of pictorial & textual MR representation (20:80)
- Pictorial MRs:
  - elicit more varied, better rated texts
  - cause less lexical priming
  - add some noise (not all attributes always realized)
- Quality control
- More references collected for 1 MR


```
name [Loch Fyne],
eatType[restaurant],
food[Japanese],
price[cheap],
kid-friendly[yes]
```

Novikova et al. SIGDIAL 2017
[ACL W17-5525]

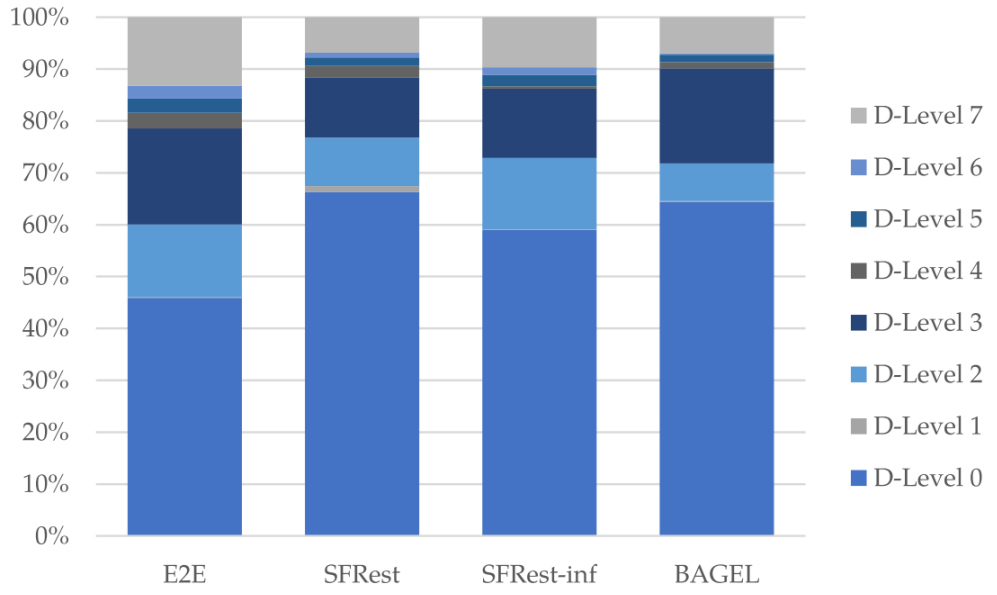
HERIOT WATT UNIVERSITY

E2E Dataset comparison

- vs. BAGEL & SFRest:
  - Lexical richness
    - **higher lexical diversity**
      (Mean Segmental Token-Type Ratio)
    - higher proportion of **rare words**

Delexicalized sets	E2E	SFRest	SFRest-inf	BAGEL
Distinct tokens	**2,675**	504	405	183
Lexical sophistication (LS2)	**0.600**	0.323	0.317	0.317
Type-token ratio (TTR)	0.002	0.012	0.013	**0.035**
Mean segmental TTR (MSTTR-50)	**0.663**	0.602	0.553	0.478

  - Syntactic richness
    - **more complex** sentences (D-Level)



The Vaults is an Indian restaurant.

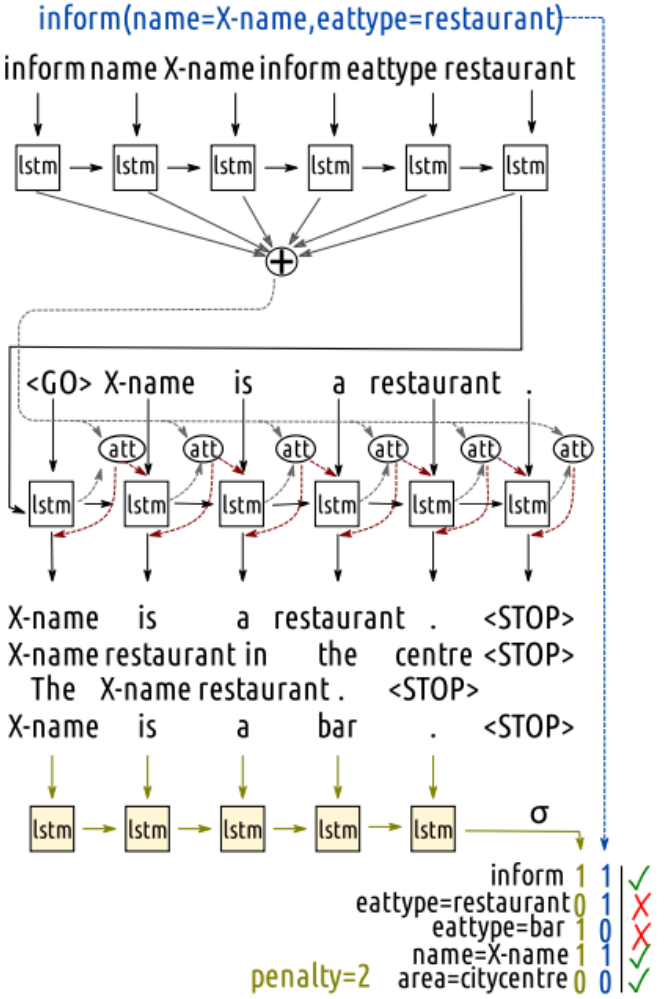
Cocum is a very expensive restaurant but the quality is great.

The coffee shop Wildwood has fairly priced food, while being in the same vicinity as the Ranch.

Serving cheap English food, as well as having a coffee shop, the Golden Palace has an average customer ranking and is located along the riverside.

Baseline model

- **TGen** (**http://bit.ly/TGen-nlg**)
- Seq2seq + attention
- **Beam reranking** by MR classification
  - any differences w.r.t. input MR penalized
- Delexicalization
  - replacing with placeholders
  - open-set attributes only (name/near)
- Strong (near SotA)



Challenges: Semantic control

- most systems attempt to realize all attributes
- template/rule-based: given by architecture – no problem
- seq2seq: attention (all) + more:
  - **beam reranking** – MR classification, heuristic aligner, attention weights
  - modifying attention (regularization)
- other data-driven:
  - **ZHAW1, ZHAW2**: semantic gates (SC-LSTM)
  - **SHEFF1**: given by architecture (realizing slots → values)

Challenges: Open vocabulary

- E2E data: name/near slots (restaurant names)
- mostly addressed by delexicalization (placeholders)
  - rule + template-based: all systems, all slots
  - data-driven: most systems, mostly name/near
- alternatives – seq2seq systems:
  - copy mechanism (**CHEN, HARV, ADAPT**)
  - sub-word units (**ZHANG**)
  - character-level seq2seq (**NLE**)

Challenges: Diversity

- data augmentation
  - to enlarge training set (SLUG)
  - for more robustness (TNT1, TNT2)
- data selection
  - using only the “most common” example: SHEFF1
  - using only more complex examples: SLUG-ALT
- diverse ensembling: HARV
- preprocessing
  - for diversity: ZHAW1, ZHAW2, ADAPT