CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

Challenges in Neural NLG

Ondřej Dušek

Institute of Formal and Applied Linguistics, Charles University

work partly done at **The Interaction Lab, Heriot-Watt University**
in collaboration with
David M. Howcroft, **Ioannis Konstas**, **Jekaterina Novikova**,
Verena Rieser & **Karin Sevegnani**

16 Oct 2019

CHARLES UNIVERSITY

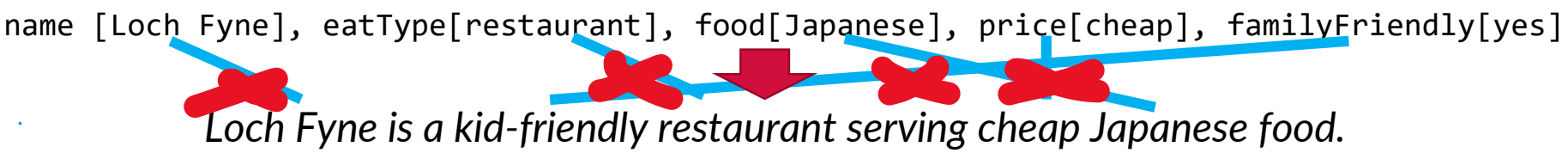
HERIOT WATT UNIVERSITY

Outline of this talk

- **E2E NLG Challenge**
  - extended results & take-away notes
- **Improving E2E data**
  - removing semantic noise
  - testing the effect on generators
- **NLG quality estimation**
  - training a system to evaluate NLG outputs
  - beyond BLEU: no references needed

E2E NLG Challenge

- Task: generating restaurant recommendations
  - simple input MR
  - no content selection expected (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?

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

E2E Dataset

- 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]
```

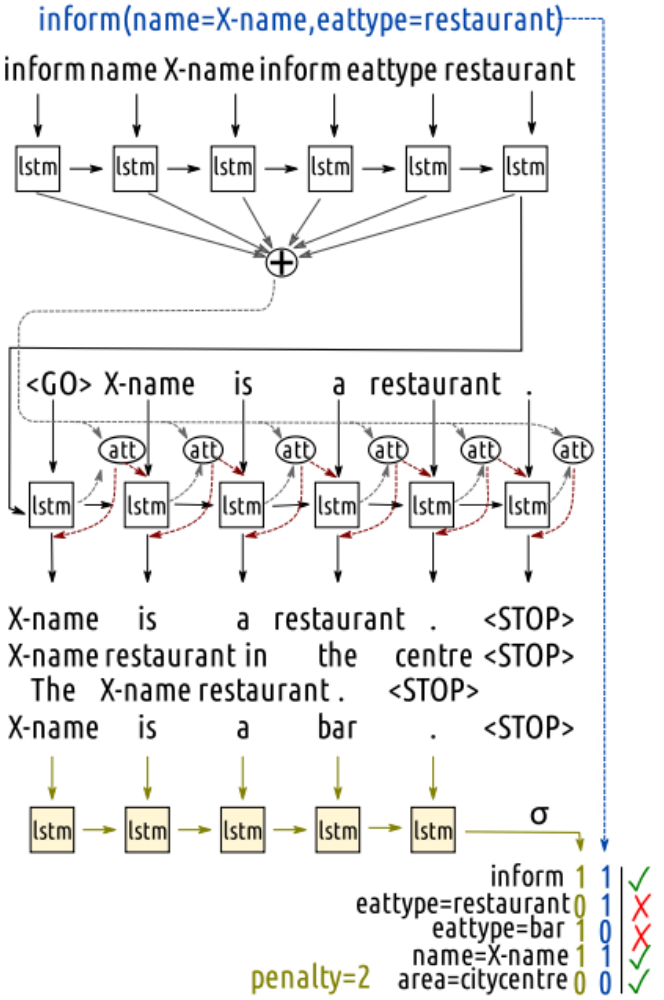
Loch Fyne is a kid-friendly restaurant serving cheap Japanese food.

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

Dušek & Jurčíček
ACL 2016
[ACL P16-2008]

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, even now!)



E2E Participants

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

- 62 systems, 17 groups (20 primary)
- Seq2seq: 12 systems + baseline
  - many variations & additions
- Other 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

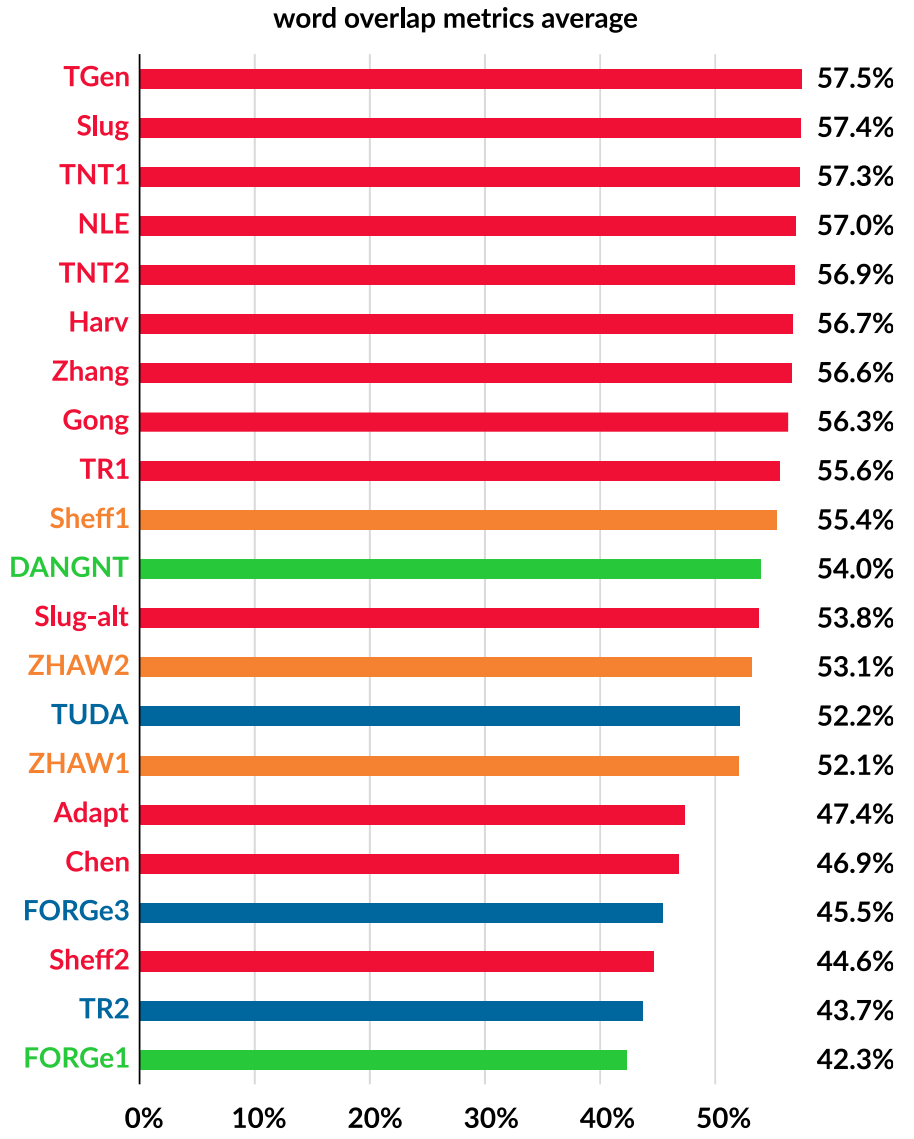
TGEN	HWU (baseline)	*seq2seq + reranking*
SLUG	UCSC Slug2Slug	*ensemble seq2seq + reranking*
SLUG-ALT	UCSC Slug2Slug	*SLUG + data selection*
TNT1	UCSC TNT-NLG	*TGEN + data augmentation*
TNT2	UCSC TNT-NLG	*TGEN + data augmentation*
ADAPT	AdaptCentre	*preprocessing step + seq2seq + copy*
CHEN	Harbin Tech (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 Challenges

- **Open vocabulary** (restaurant/landmark names)
  - delexicalization (placeholders)
    - template/rule-based: everything, not just restaurant/landmark names
  - seq2seq: also 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 (SLUG, SLUG-ALT)
  - diverse ensembling (HARV)
  - preprocessing steps (ZHAW1, ZHAW2, ADAPT)

Word-overlap metrics

- BLEU, NIST, METEOR, ROUGE, CIDEr
  - Scripts provided
    - http://bit.ly/e2e-nlg
  - Segment-level correlation vs. humans weak (<0.2)
- Baseline very strong
- Seq2seq systems on top
  - very close competition
  - not all seq2seq systems – some very low
- Handcrafted systems at the bottom



Output Complexity

- **Syntactic complexity**: wide range
  - low: seq2seq systems (except SLUG-ALT)
    - more plain sentences, very few complex ones compared to other systems & human references
  - high: FORGE3, TR2, ZHAW1, ZHAW2, SHEFF1, FORGE1, SLUG-ALT
    - more complex sentences than human average
- **Rare words ratio** lower than human refs
  - high: ADAPT, FORGE1, TR2 (still not matching human)
- **Output length**: wide range
  - high: handcrafted systems (longer than human)
  - low: low-performing seq2seq – very short outputs (CHEN, SHEFF2)

Output Diversity

- **Distinct tokens/n-grams**: all systems worse than humans
  - high: template mining (TR2, FORGE3),
    diversity-attempting (ZHAW1, ADAPT, SLUG-ALT)
  - low: plain seq2seq
    fully handcrafted (DANGNT, TUDA)
- **Entropy**: all systems worse than humans
  - similar system ranking to tokens/n-grams
  - bigram conditional entropy – huge gap

Bigram cond. entropy	
test set all	2.92
test set rand	2.70
TR2	2.60
ADAPT	2.09
FORGE3	1.66
SLUG-ALT	1.55
HARV	1.45
ZHAW1	1.44
TNT2	1.39
NLE	1.37
TNT1	1.37
SHEFF1	1.33
TGEN	1.32
ZHAW2	1.32
TR1	1.30
FORGE1	1.29
ZHANG	1.26
CHEN	1.17
SLUG	1.13
SHEFF2	1.10
DANGNT	1.06
GONG	0.91
TUDA	0.71

Semantic Accuracy

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

- Measured by pattern matching script
- Handcrafted systems score mostly high
  - template mining bad (**TR2**, **FORGE3**)
- Some data-driven systems also good
  - **SHEFF1**, **GONG**, **SLUG**
  - all have very strong semantic control
- Seq2seq without reranking very low
  - **CHEN**, **SHEFF2**, **ZHANG**, **ADAPT**, **TR1**
- Missing information most common
- Diversity may hurt accuracy
  - **HARV**, **FORGE3**, **TR2**, **ADAPT**
  - not **ZHAW1**, **ZHAW2**, **SLUG-ALT**

System	Insts OK	Add info	Miss info	Add+Miss	SER
TUDA	100%	0%	0%	0%	0.00%
SHEFF1	93%	0%	5%	2%	1.08%
GONG	92%	4%	2%	2%	1.13%
FORGE1	92%	0%	8%	0%	1.22%
SLUG	91%	1%	4%	4%	1.26%
DANGNT	88%	0%	12%	0%	1.75%
TGEN	79%	3%	16%	2%	3.56%
SLUG-ALT	78%	4%	9%	9%	3.56%
ZHAW2	76%	3%	20%	1%	3.68%
TNT1	73%	1%	22%	4%	4.92%
TNT2	71%	1%	28%	1%	6.04%
ZHAW1	70%	3%	25%	2%	5.12%
TR2	66%	6%	23%	5%	5.45%
NLE	63%	3%	24%	10%	6.20%
HARV	54%	2%	30%	14%	10.43%
ADAPT	50%	3%	36%	10%	12.48%
TR1	48%	0%	52%	0%	13.83%
FORGE3	41%	0%	55%	3%	10.41%
ZHANG	27%	0%	73%	0%	14.80%
CHEN	11%	0%	88%	1%	23.53%
SHEFF2	5%	0%	88%	6%	27.94%

Human evaluation

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

CHARLES UNIVERSITY

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 architectures
  - 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

- NB: 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
- **Complexity & Diversity** – hand-engineered systems seem better
  - hand-engineered: direct control
  - options for seq2seq: diverse ensembling, sampling…
  - diversity might hurt naturalness
- **Best method**: rule-based or seq2seq with reranking

Mismatched semantics in E2E

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

- E2E data is noisy, but has multiple references per MR (ø=8.1)
  - Using the SER script on the data: SER=16.37%
- Challenge system outputs are noisy
  - some systems less, some systems more
  - what happens if we clean the data?
- We can apply the SER script & assign matching MRs
  - cleans the data & doesn't change human-produced texts

```
name[Cotto], eatType[coffee shop], food[English], priceRange[less than £20],
customer rating[low], area[riverside], near[The Portland Arms]
```

Cotto is a coffee shop that serves English food in the city centre.
They are located near the Portland Arms and are low rated. (missing price)

Mismatched semantics in E2E

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

- E2E data is noisy, but has multiple references per MR (ø=8.1)
  - Using the SER script on the data: SER=16.37%
- Challenge system outputs are noisy
  - some systems less, some systems more
  - what happens if we clean the data?
- We can apply the SER script & assign matching MRs
  - cleans the data & doesn't change human-produced texts

```
name[Cotto], eatType[coffee shop], food[English], priceRange[less than £20],
customer rating[low], area[riverside city centre], near[The Portland Arms]
```

Cotto is a coffee shop that serves English food in the city centre.
They are located near the Portland Arms and are low rated. (missing price)

E2E Data Cleaning

- Ensure train-test split → cleaned smaller than original
  - needed to remove fixed MRs from train/dev that overlapped test set
  - allows testing on original test set
- More distinct MRs
- Fewer references per MR
  - → cleaned test set more challenging than original
    - numbers on test sets not comparable
- Manual check of SER script – 4.2% SER
  - not perfect, but much lower than data original

	Set	MRs	References
Original Data	Train	4,862	42,061
	Dev	547	4,672
	Test	630	4,693
Cleaned Data	Train	8,362	33,525
	Dev	1,132	4,299
	Test	1,358	4,693

Training on Cleaned Data

- Plain Seq2seq
- TGen (seq2seq + reranking)
- BLEU – not much change
- **SER lowered by 97%** (for TGen)!
  - Seq2seq – 94% error reduction
- Main improvement comes from fixing missed slots
  - just fixing added has little effect

Train	System	BLEU	NIST	% Instances with Added slots	Missed slots	Wrong values	% SER
Original	Seq2seq	63.37	7.71	0.06	15.77	0.11	15.94
Original	TGen	66.41	8.55	0.14	4.11	0.03	4.27
Cleaned	Seq2seq	65.87	8.64	0.20	0.56	0.21	0.97
Cleaned	TGen	66.24	8.68	0.10	0.02	0.00	0.12
Cleaned missing	Seq2seq	66.28	8.52	0.14	2.26	0.22	2.61
Cleaned missing	TGen	67.00	8.68	0.06	0.44	0.03	0.53
Cleaned added	Seq2seq	64.40	7.96	0.01	13.08	0.00	13.09
Cleaned added	TGen	66.23	8.55	0.04	3.04	0.00	3.09

Manual analysis

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

- TGen only, on a sample
  - just error analysis, done by us
  - not user rating
- Confirms automatic results
  - the more you clean, the better
  - disfluencies are *very* slight
- SER script very accurate on system outputs (99.9% correct)
  - more reliable than on the training data
    - outputs have more frequent phrasing

TGen / trained on	Add	Miss	Wrong	Disfluencies
Original	0	22	0	14
Cleaned added	0	23	0	14
Cleaned missing	0	1	0	2
Cleaned	0	0	0	5

NLG Quality Estimation

- Given NLG output, check if it's good or not (give rating)
  - **without using reference texts**
  - can be used at runtime: should we trigger a fallback?
- Given more outputs, which is the best?
  - selecting from n-best list
- System development: can QE be better than BLEU?
- Trained from few manual ratings

MR: `inform_only_match(name='hotel drisco', area='pacific heights')`
NLG output: the only match i have for you is the hotel drisco in the pacific heights area.

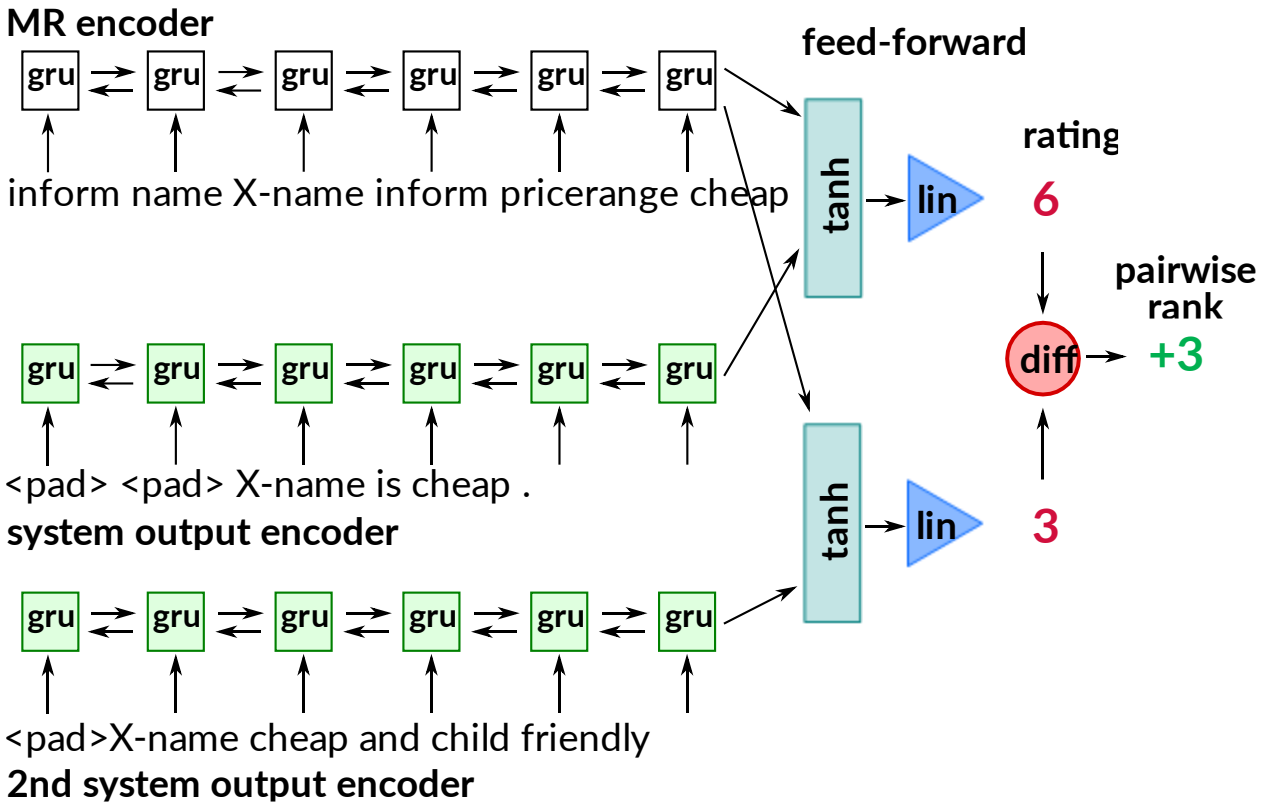
Rating:
4 (on a 1-6 scale)

MR: `inform(name='The Cricketers', eatType='coffee shop', rating=high, familyFriendly=yes, near='Café Sicilia')`
NLG 1: The Cricketers is a children friendly coffee shop near Café Sicilia with a high customer rating .
NLG 2: The Cricketers can be found near the Café Sicilia. Customers give this coffee shop a high rating. It's family friendly.

Rank:
better
worse

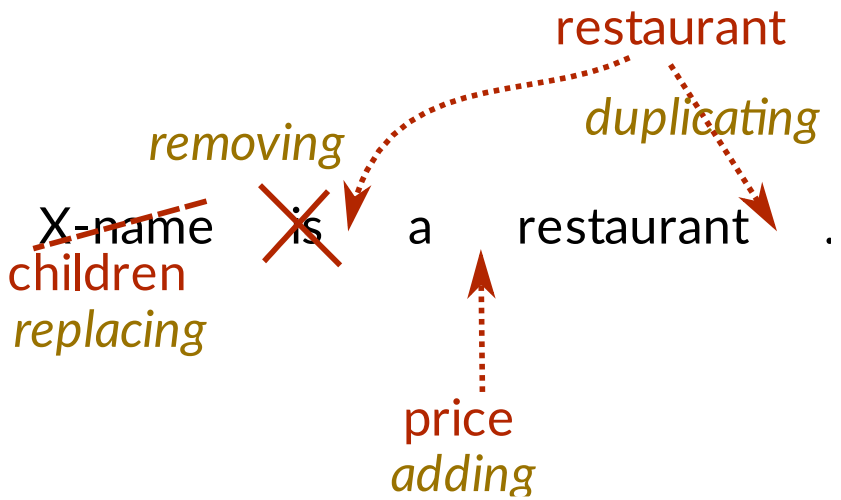
Our Model

- Rating: dual encoder
  - MR encoder
  - NLG output encoder
  - fully connected + linear
  - trained by **squared error**
- Ranking extension:
  - 2nd copy NLG output encoder + fully connected + linear
    - shared weights
  - trained by **hinge rank loss**
    - on difference from 2 ratings
- Can learn ranking & rating jointly
  - training instances mixed & losses masked



Synthetic Data

- Adding more training instances
  - introducing artificial errors
  - randomly:*
    - removing words
    - replacing words by random ones
    - duplicating words
    - inserting random words
- Rating instances
  - lower rating by 1 per error (+ 6 → 4)
- Ranking instances
  - different # of errors introduced: rank fewer errors better
  - can be used when system is trained for ratings, too!



* articles and punctuation are dispreferred

Results: Rating

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

- Small 1-6 Likert-scale data (2,460 instances)
  - 3 systems, 3 datasets (hotels & restaurants)
  - 5-fold cross-validation
- Much better correlations than BLEU et al.
  - despite not needing references
  - synthetic data help a lot
    - statistically significant
  - absolute value still not great
    - noise in human data?
- MAE/RMSE not so great
  - synthetic data don't help for that

System	Pearson	Spearman	MAE	RMSE
Constant	-	-	1.013	1.233
BLEU (needs human references)	0.074	0.061	2.264	2.731
Our base	0.253	0.252	**0.917**	**1.221**
+ synthetic rating instances	0.332	0.308	0.924	1.241
+ synthetic ranking instances	0.347	**0.320**	0.936	1.261
+ synthetic from human references	**0.369**	0.295	0.925	1.250

Results: Ranking

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

- Using E2E human ranking data (quality) – 15,001 instances
  - 21 systems, 1 domain
  - 5-way ranking converted to pairwise, leaving out ties
  - 8:1:1 train-dev-test split, no MR overlap
- Much better than random
- Synthetic ranking instances help
  - +4% absolute, statistically significant
- Training on both datasets doesn't help
  - different text style, different systems

System	P@1/Acc
Random	0.500
Our base	0.708
+ synthetic ranking instances	0.732
+ synthetic from human references	**0.740**

Conclusions

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

- Neural generators can work really well
  - but handcrafted ones still work nicely for limited domains
- If you want diversity, you need a lot of data either way
  - seq2seq + diversity measures / template mining
  - need to be careful about accuracy!
- Neural generators don't like noisy training data
  - outputs are accurate if training data is accurate
  - less + clean is better than more + noisy
- Trained quality estimation can do better than BLEU
  - synthetic ranking instances help
  - still has a problem with domain/system generalization
- Future work: using pretrained LMs

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

Thanks

- Get these slides:
  http://bit.ly/dusek-nlg-2019
- Contact me:
  odusek@ufal.mff.cuni.cz
  http://bit.ly/odusek
  @tuetschek

E2E Challenge: Computer Speech & Language 59, arXiv: 1901.07931, http://bit.ly/e2e-nlg

Cleaning E2E: INLG 2019, arXiv coming soon, http://bit.ly/clean-e2e

NLG QE: INLG 2019, arXiv: 1910.04731, http://bit.ly/ratpred

CHARLES
UNIVERSITY

HERIOT
WATT
UNIVERSITY

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

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)

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

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

Complexity & Diversity

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

% Level0-2		% Level6-7		LS2		MSTTR-50		Avg. length	
♡GONG	82.68	◇SHEFF1	41.27	*test set all*	0.43	*test set rand*	0.62	♠TUDA	31.02
♡TNT2	79.64	♣FORGE1	33.66	*test set rand*	0.36	♠TR2	0.62	♠TR2	27.48
♡SLUG	78.08	♡SLUG-ALT	30.49	♡ADAPT	0.33	♡ADAPT	0.61	♣FORGE1	26.88
♡TNT1	72.18	◇ZHAW1	26.00	♣FORGE1	0.30	♣FORGE1	0.59	◇ZHAW2	26.58
♡ZHANG	70.83	♠TR2	21.07	♠TR2	0.29	◇ZHAW1	0.58	♡TNT1	26.37
♣DANGNT	66.95	◇ZHAW2	19.03	♡HARV	0.27	*test set all*	0.58	◇ZHAW1	26.16
♡TGEN	65.12	♠FORGE3	18.51	♡TNT1	0.26	◇ZHAW2	0.57	♡TNT2	25.49
♡HARV	64.63	*test set rand*	17.46	♡CHEN	0.25	♠FORGE3	0.56	♡GONG	25.41
♡TR1	64.28	♡GONG	16.90	♡NLE	0.25	♠TUDA	0.55	♣DANGNT	24.85
♠FORGE3	62.62	*test set all*	16.48	♡SHEFF2	0.25	♣DANGNT	0.54	♡ADAPT	24.47
♡ADAPT	62.48	♡SLUG	11.39	◇SHEFF1	0.24	♡SLUG-ALT	0.54	♡SLUG-ALT	24.47
♣FORGE1	61.13	♡NLE	11.12	♡TNT2	0.23	♡SLUG	0.52	*test set rand*	24.39
◇ZHAW1	58.91	♠TUDA	10.48	♡TGEN	0.22	♡TNT1	0.52	♡TGEN	24.04
♡NLE	58.24	♡ADAPT	10.28	♣DANGNT	0.21	◇SHEFF1	0.52	*test set all*	23.96
test set rand	58.16	♡TNT1	9.55	♠TUDA	0.21	♡NLE	0.52	♡SLUG	23.76
test set all	57.97	♡TGEN	9.02	♡TR1	0.20	♡TGEN	0.52	♠FORGE3	23.49
♠TUDA	57.66	♣DANGNT	8.91	♡ZHANG	0.20	♡TNT2	0.51	♡NLE	23.40
♠TR2	57.36	♡TR1	8.13	♡SLUG	0.20	♡HARV	0.51	♡HARV	23.22
♡CHEN	54.35	♡HARV	8.12	♡GONG	0.20	♡TR1	0.50	◇SHEFF1	22.75
♡SHEFF2	52.98	♡ZHANG	5.27	♠FORGE3	0.20	♡GONG	0.50	♡TR1	22.43
◇ZHAW2	52.63	♡TNT2	5.22	♡SLUG-ALT	0.19	♡ZHANG	0.47	♡ZHANG	20.71
♡SLUG-ALT	35.12	♡CHEN	4.40	◇ZHAW2	0.17	♡CHEN	0.43	♡SHEFF2	17.18
◇SHEFF1	26.19	♡SHEFF2	2.08	◇ZHAW1	0.17	♡SHEFF2	0.43	♡CHEN	16.32

Diversity

CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

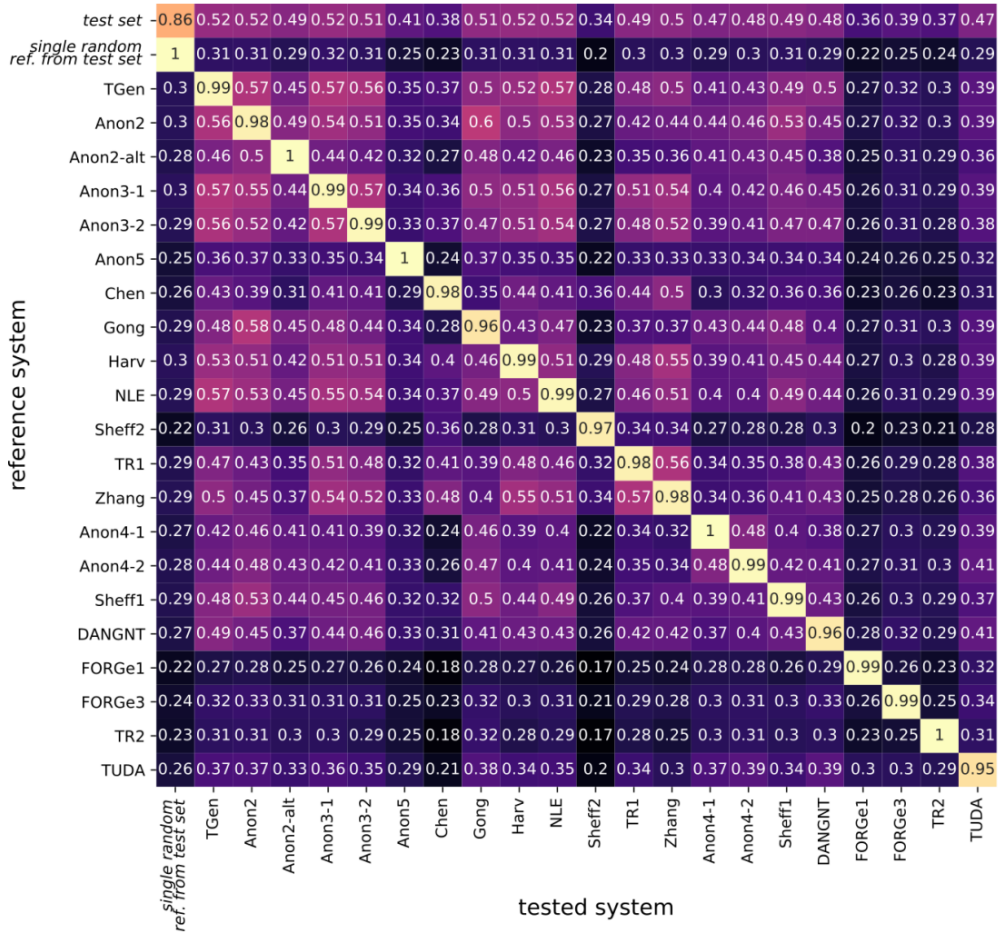
Distinct tokens		Distinct trigrams		% Unique trigrams		Entropy tokens		Cond. entropy bigrams	
test set all	1079	*test set all*	16797	*test set rand*	69.13	*test set all*	6.40	*test set all*	2.92
test set rand	542	*test set rand*	5166	♡ADAPT	66.61	*test set rand*	6.37	*test set rand*	2.70
♡ADAPT	455	♠TR2	4687	♠TR2	60.44	♠TR2	6.24	♠TR2	2.60
♠TR2	399	♡ADAPT	3567	*test set all*	44.66	♡ADAPT	6.18	♡ADAPT	2.09
◇ZHAW1	136	◇ZHAW1	969	◇ZHAW1	24.97	♠FORGE3	5.74	♠FORGE3	1.66
♠FORGE3	124	♠FORGE3	896	♡HARV	21.88	◇ZHAW1	5.71	♡SLUG-ALT	1.55
◇ZHAW2	102	♡SLUG-ALT	855	♡TNT1	21.34	◇ZHAW2	5.65	♡HARV	1.45
♡HARV	93	♡HARV	777	♡NLE	18.75	♡SLUG-ALT	5.57	◇ZHAW1	1.44
♡TNT1	89	◇ZHAW2	716	◇ZHAW2	18.72	♣FORGE1	5.55	♡TNT2	1.39
♣FORGE1	88	♡TNT1	703	♡SLUG-ALT	18.13	♡HARV	5.50	♡NLE	1.37
♡SLUG-ALT	88	♡TNT2	634	♡CHEN	17.92	◇SHEFF1	5.43	♡TNT1	1.37
♡TNT2	86	♡NLE	608	♡ZHANG	17.81	♡NLE	5.43	◇SHEFF1	1.33
♡TGEN	83	♡TGEN	597	◇SHEFF1	16.44	♡TGEN	5.41	♡TGEN	1.32
♡NLE	81	◇SHEFF1	578	♡SLUG	15.58	♡TNT1	5.37	◇ZHAW2	1.32
♡ZHANG	76	♣FORGE1	549	♠FORGE3	13.50	♡SLUG	5.35	♡TR1	1.30
♡TR1	75	♡ZHANG	511	♡TGEN	13.23	♡TNT2	5.34	♣FORGE1	1.29
♡SLUG	74	♡SLUG	507	♡TNT2	12.93	♣DANGNT	5.29	♡ZHANG	1.26
♡CHEN	73	♡CHEN	480	♣FORGE1	12.39	♠TUDA	5.25	♡CHEN	1.17
◇SHEFF1	72	♡TR1	464	♡TR1	10.78	♡TR1	5.24	♡SLUG	1.13
♣DANGNT	61	♣DANGNT	301	♡GONG	7.30	♡ZHANG	5.21	♡SHEFF2	1.10
♡SHEFF2	59	♡SHEFF2	262	♡SHEFF2	4.96	♡GONG	5.19	♣DANGNT	1.06
♡GONG	58	♡GONG	233	♣DANGNT	0.00	♡CHEN	5.09	♡GONG	0.91
♠TUDA	57	♠TUDA	143	♠TUDA	0.00	♡SHEFF2	4.76	♠TUDA	0.71

Output diversity

- MSTTR-50: only the highest match humans
  - high: handcrafted systems (TR2, FORGE1, FORGE3, TUDA)
    systems aiming at diversity (ZHAW1, ZHAW2, ADAPT, SLUG-ALT)
  - low: plain seq2seq
  - may be skewed by output length

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

- 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]

- 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]

CHARLES UNIVERSITY

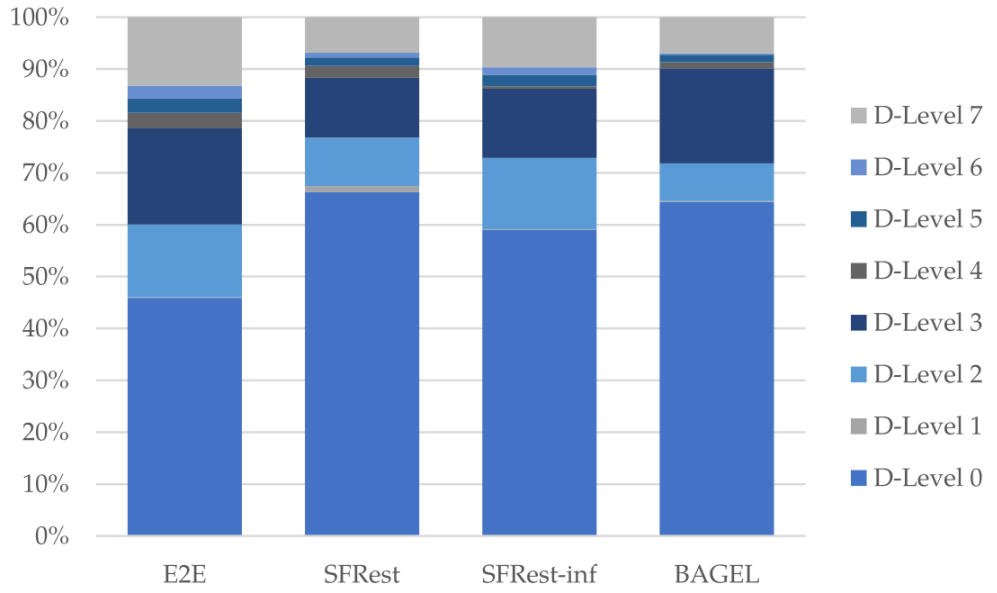
HERIOT WATT UNIVERSITY

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.