CHARLES UNIVERSITY

HERIOT WATT UNIVERSITY

Automatic Quality Estimation for Natural Language Generation: Ranting (Jointly Rating and Ranking)

Ondřej Dušek, Karin Sevegnani, Ioannis Konstas & Verena Rieser

Charles University, Prague
Heriot-Watt University, Edinburgh

INLG, Tokyo, 31 Oct 2019

Our Task(s)

- **Quality estimation**: checking NLG output quality
  - just given input MR & NLG system output
  - **no human reference texts** for the NLG output
  - **supervised training** from a few human-annotated instances
  - well-established for MT, not so much in data-to-text NLG
- **Rating**: Given NLG output, check if it's good or not (scale 1-6)
- **Ranking**: Given more NLG outputs, which one is the best?

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 . ← better
NLG 2: The Cricketers can be found near the Café Sicilia. Customers give this coffee shop a high rating. It's family friendly. ← worse

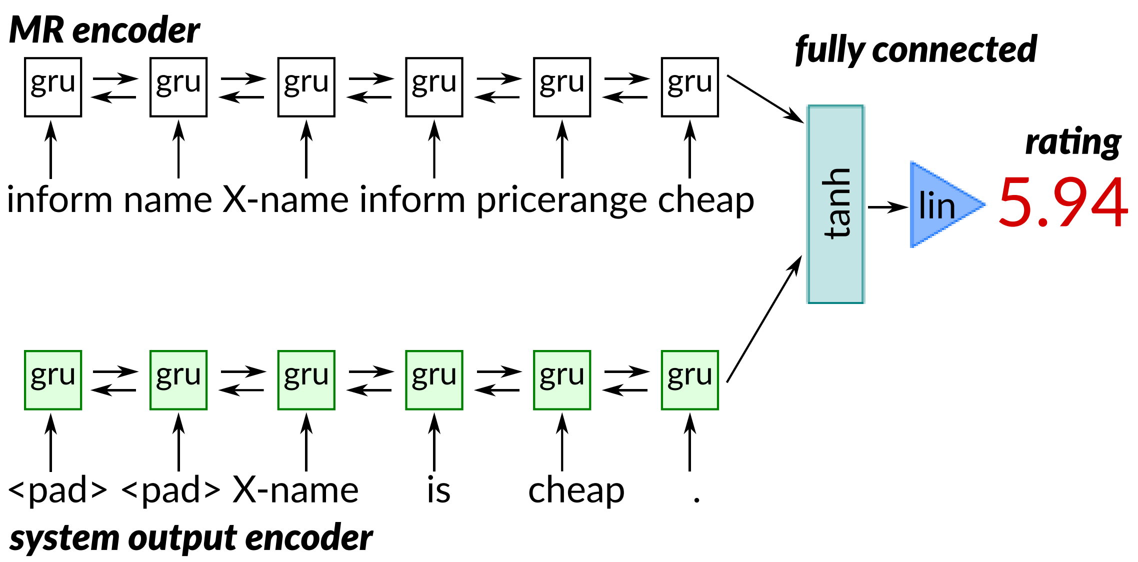
Rank:

Why Quality Estimation?

- BLEU et al. don’t work very well – can we be better?
  - evaluating via correlation with humans
- We can do without human references – wider usage:
  - Evaluation, tuning (same as BLEU)
  - Tuning (same as BLEU)
  - Inference – improving running NLG systems
- Inference time use:
  - **for rating**: don’t show outputs rated below a threshold
    - use a backoff or humans
  - **ranking**: select best system output from an n-best list

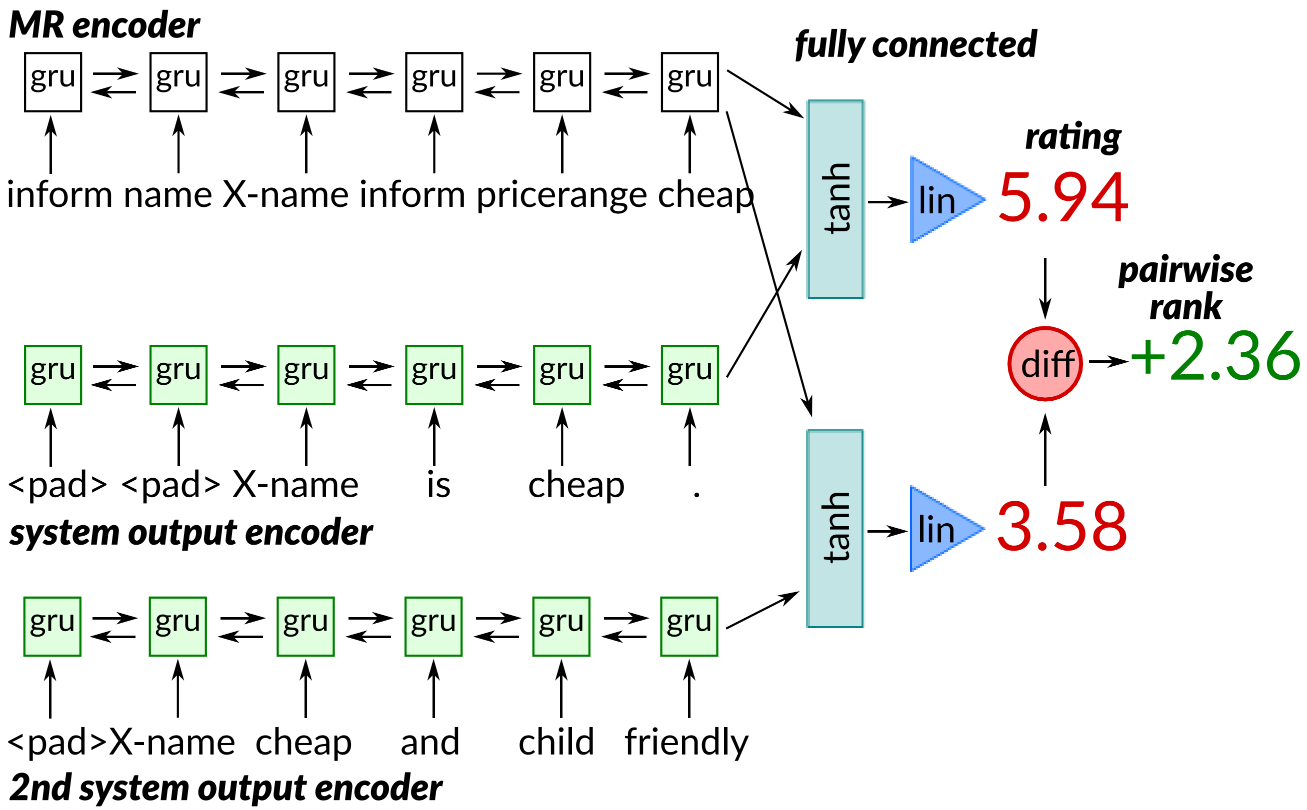
Old Model (Dušek, Novikova & Rieser, 2017)

- Ratings only
- Dual-encoder
  - MR encoder
  - NLG output encoder
  - fully connected + linear
  - trained by **squared error**
- Final score is rounded



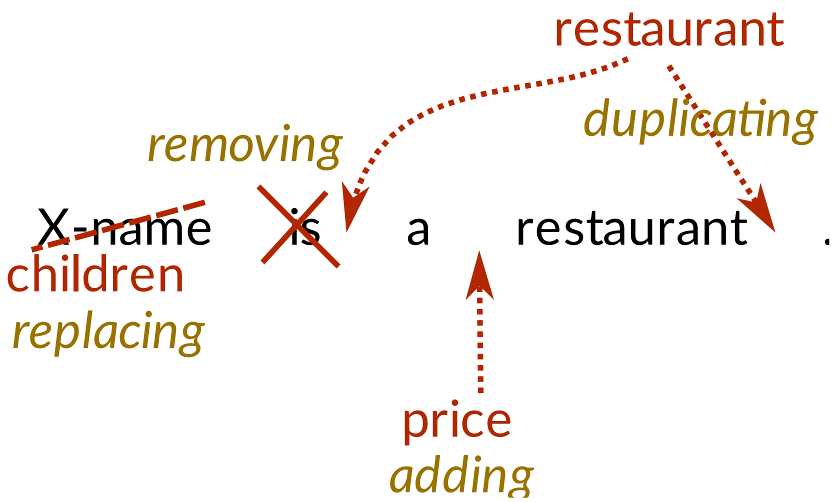
Our Model

- 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 (Dušek, Novikova & Rieser, 2017)

- Adding more training instances
  - introducing artificial errors
  - randomly:*
    - removing words
    - replacing words by random ones
    - duplicating words
    - inserting random words
- For rating data:
  - lower the rating by 1 for each error (with 6 → 4)
- This can be applied to NLG systems' training data, too
  - assume 6 (maximum) as original instances' rating



* articles and punctuation are dispreferred

Synthetic Ranking Pairs

- Different #'s of errors introduced to the same NLG output
- Fewer errors should rank better
- Ranking pairs are useful when the system is trained to rate, too!

restaurant

X-name ~~serves~~ Chinese food . ← *1 error* — **Rank:** better

food, cheaply

X-name serves Chinese food . ← *2 errors* — worse

Results: Rating

- 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
  - correlation of 0.37 still not ideal
    - noise in human data?
- absolute differences (MAE/RMSE) not so great

(Novikova et al., EMNLP 2017)
https://aclweb.org/anthology/D17-1238

System	Pearson	Spearman	MAE	RMSE
Constant	-	-	1.013	1.233
BLEU (needs human references)	0.074	0.061	2.264	2.731
Our previous (Dušek et al., 2017)	0.330	0.287	**0.909**	**1.208**
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 systems’ training data	**0.369**	0.295	0.925	1.250

Results: Ranking

(Dušek et al., CS&L 59)
https://arxiv.org/abs/1901.07931

- 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
- Our system is much better than random in pairwise ranking accuracy
- 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 systems' training data	**0.740**

Conclusions

- Trained quality estimation can do much better than BLEU & co.
  - Pearson correlation with humans 0.37 vs. ~0.06-0.10
  - synthetic ranking instances help
- The results so far aren't ideal (we want more than 0.37/74%)
- Domain/system generalization is still a problem
- Future work:
  - improving model
  - using pretrained LMs
  - obtaining "cleaner" user scores
  - more realistic synthetic errors
  - influence of error type on user ratings

Thanks

- Code & link to data + paper:
  http://bit.ly/ratpred
- Contact me:

  odusek@ufal.mff.cuni.cz
  http://bit.ly/odusek
  @tuetschek

Paper links:	this paper:	arXiv: 1910.04731
	previous model:	arXiv: 1708.01759
	datasets used:	ACL D17-1238, arXiv:1901.07931