Referenceless Quality Estimation for Natural Language Generation

Ondřej Dušek, Jekaterina Novikova & Verena Rieser
Interaction Lab, Heriot-Watt University, Edinburgh

ICML-LGNL, Sydney, August 2017

HERIOT WATT UNIVERSITY

Quality Estimation for NLG

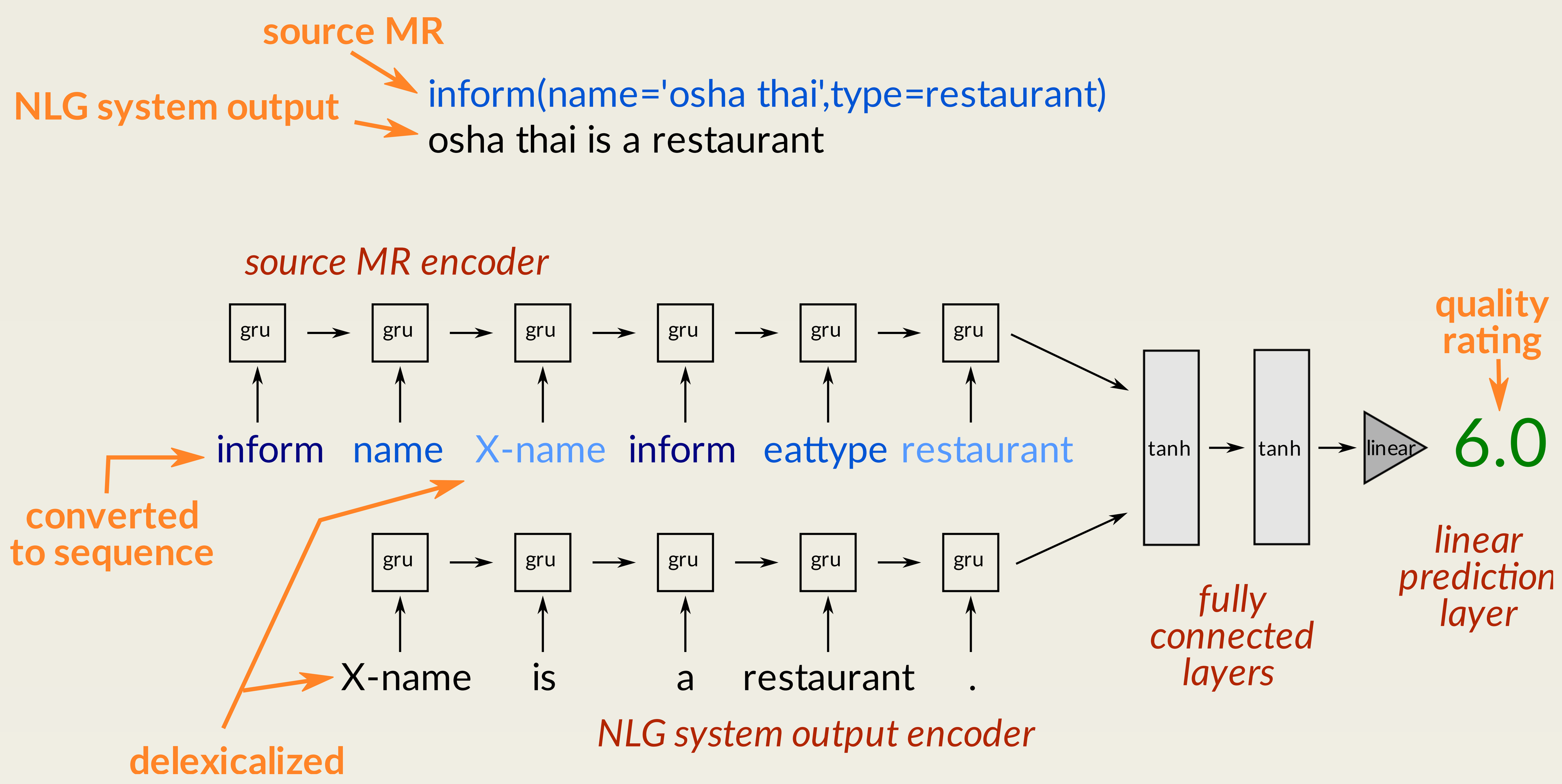
- estimate **NLG system output quality** by **comparing with source MR only** – **no reference texts** needed
- useful **for system development**: word-overlap metrics such as BLEU unreliable + need costly references
- useful **at runtime**: reranking, triggering fallback strategies

	Instance	Humans	BLEU	METEOR	ROUGE	CIDEr	Our system (S4)
MR	inform(name=‘la ciccia’,area=‘bernal heights’,price_range=moderate)	5.51		3	3.5	2	4.5
System output	la ciccia, is in the bernal heights area with a moderate price range.		*0.000*	*0.371*	*0.542*	*2.117*	
Reference	*la ciccia is a moderate price restaurant in bernal heights*						
MR	inform(name=‘intercontinental san francisco’,price_range=‘pricey’)	2	4.5	3	5.5	2	5
System output	the intercontinental san francisco is in the pricey price range.		*0.707*	*0.433*	*0.875*	*2.318*	
Reference	*sure, the intercontinental san francisco is in the pricey range.*						

Our NLG Quality Estimation Model

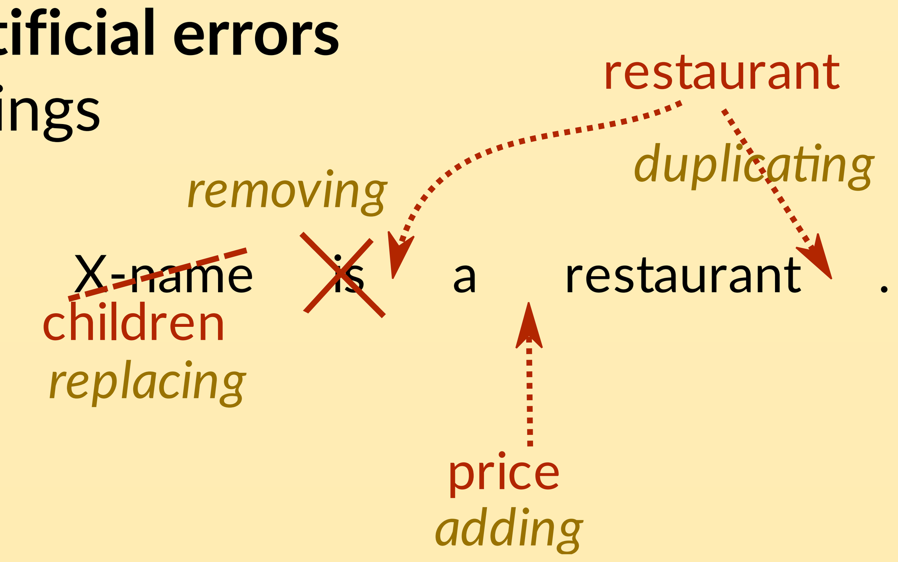
1) RNN **GRU encoders** for source MR + NLG system output to be rated
2) fully connected ***tanh* layers**
3) final layer – **linear**, predicting rating as a floating point number

- trained by **minimising mean square error** against human-assigned ratings
- delexicalization to fight data sparsity



Our Dataset

- **outputs of 3 NLG systems on 3 datasets**
  - TGen & LOLS & RNNLG
  - BAGEL & SFHot & SFRest
- **CrowdFlower** used to obtain human ratings
  - **overall quality** rating on a **1–6 Likert scale**
  - 3+ ratings per system output
  - using medians for consistency
- 2,460 instances total
- **synthesising additional data**:
  - a) introducing **artificial errors** & lowering ratings



  - b) additional **human references** from source NLG datasets (with “perfect” ratings)
    - up to 78k synthesised instances

Setup	Pearson	Spearman	MAE	RMSE
Constant	-	-	1.013	1.233
BLEU*	0.074	0.061	2.264	2.731
METEOR*	0.095	0.099	1.820	2.129
ROUGE-L*	0.079	0.072	1.312	1.674
CIDEr*	0.061	0.058	2.606	2.935
S1: Base system	0.273	0.260	0.948	1.258
S2: + artifical errors in training system outputs	0.283	0.268	0.948	1.273
S3: + human references for training MRs (& errors)	0.278	0.261	0.930	1.257
S4: + human refs from source NLG training sets (& errors)	**0.330**	0.274	0.914	1.226
S5: + human references for test MRs (& errors)*	**0.331**	0.265	0.937	1.245
S6: + human refs from whole source NLG sets (& errors)*	**0.354**	**0.287**	0.909	1.208

bold = significantly better correlation than S1
* = using human reference texts for test MRs (i.e. not strictly referenceless)

Experiments & Results

- 5-fold cross-validation
- always **better correlations than metrics**
  - lower than MT (less data & harder)
  - 21% improvement with synthetic data
- with synthetic data: **better MAE/RMSE than constant baseline**
- cross-domain & cross-system performance poor, but small amounts of in-set data help greatly

Conclusions

- 1st quality estimation system for NLG
- no need for references, better segment-level correlations than word-overlap metrics
- improvements over constant baseline suggest occasional large errors
- code available at:
  https://github.com/tuetschek/ratpred
- **future work:** better networks, better error synthesis, more data (**E2E NLG challenge**), post-edits prediction