Referenceless Quality Estimation for Natural Language Generation

Ondřej Dušek, Jekaterina Novikova & Verena Rieser

Interaction Lab
Heriot-Watt University
Edinburgh, Scotland, UK

ICML/LGNNL, Sydney, 10 August 2017
Quality Estimation for NLG

Task
• estimate NLG system output quality by comparing with input MR only
• no human-authored reference texts needed

Motivation
• human references are costly
• word-overlap metrics (e.g. BLEU) have low correlation with human ratings

Usage
• NLG system development + runtime: reranking, triggering fallback
Our Model & Data

Model

- Neural network, trained on human-assigned ratings
- 2 RNN encoders (for MR & system output) + further layers
- output: float

Data

- crowdsourced ratings for 3 real NLG systems’ outputs on 3 datasets
- quality 1–6 Likert scale
- synthesising additional data:
  a) artificial errors
  b) using original human references from source datasets
Results

• up to 0.35 Pearson correlation with human ratings
  • synthetic data helps (21% correlation increase)

• Up to 6x better correlation than BLEU/ROUGE/METEOR/CIDEr
  • Worse than similar experiments in MT (less data & harder)

• Better than constant baseline

• Cross-domain & cross-system performance poor
  • but in-set data helps a lot
Thanks

• Come see my poster!

• Download my code: http://bit.ly/ratpred

• Contact me: o.dusek@hw.ac.uk