



RankME: Reliable Human Ratings for Natural Language Generation

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

Introduction

- NLG human ratings are often inconsistent
- Prev. research: attributed to individual preferences
- Our research: **Can they be improved by experimental design?**

Challenges in NLG human ratings

- **Consistency** – depends on design
- **Magnitude Estimation (ME):** continuous scales + reference value
- **System comparison** – direct is better
- **RankME** – direct relative ranking of multiple systems
- **Distinct criteria** (informativeness vs. naturalness)
 - ratings often correlated: **collect separately**

Conclusions

- RankME increases ICC
- **Separate collection lowers correlation** between criteria
- Overall system ranks: **TrueSkill efficient**
 - RankME + TrueSkill same ranks as RankME
 - allows fewer comparisons with many systems
- See the paper & data & FigureEight designs: <https://github.com/jeknov/RankME/>

Relative Ranking

Plain ME (joint criteria)

Continuous Scales

Likert Scales (joint criteria)

Experiments

- **3 systems** (seq2seq-based), E2E NLG dataset (restaurants)
- Slug2Slug + TGen similar BLEU, Sheffield NLP lower BLEU
- 100 outputs, 3 judges each
- Comparing:
 - Likert scales / Plain ME / RankME
 - criteria collected jointly / separately
- Measuring:
 - **correlation** between criteria
 - **inter-annotator agreement (ICC)**
 - **overall sys. rankings** (separate coll. only)

Spearman correlation between criteria

		Joint	Separate
	naturalness		
Likert	quality	0.54*	-0.01
Plain ME	quality	0.44*	-0.03
RankME	quality	0.28*	-0.04

		Joint	Separate
	informativeness		
Likert	quality	0.00	0.54*
Plain ME	quality	0.48*	0.71*
RankME	quality	0.55*	0.74*

ICC (inter-annotator agreement)

Method	Rating	Joint	Separate
Likert	naturalness	0.07	0.12
	quality	0.02	0.41*
Plain ME	informativeness	0.93*	0.78*
	naturalness	-0.03	0.27*
RankME	quality	0.22*	0.60*
	informativeness	0.59*	0.79*
RankME	naturalness	0.11	0.42*
	quality	0.10	0.68*
RankME	informativeness	0.72*	0.82*

Resulting system rankings (separate collection)

Ranking	Rating criterion & method
1. Slug2Slug	Plain ME informativeness
2. TGen	RankME quality
3. Sheffield NLP	TrueSkill quality added information

1. TGen	missing information
2. Slug2Slug	
3. Sheffield NLP	

1.-2. Slug2Slug	Plain ME quality
+ TGen	RankME informativeness
3. Sheffield NLP	Likert quality
	Likert informativeness

1.-2. Slug2Slug	Likert naturalness
+ Sheffield NLP	
3. TGen	

1.-3. Slug2Slug	Plain ME naturalness
+ TGen	RankME naturalness
+ Sheffield NLP	TrueSkill naturalness