WebNLG Challenge: Human Evaluation Results

Anastasia Shimorina
Claire Gardent, Shashi Narayan, Laura Perez-Beltrachini

15 January 2018

Introduction

This report presents the human evaluation results for the WebNLGChallenge which was held in 2017. The automatic evaluation results can be found in [Gardent et al., 2017a]. In this report, we describe human evaluation design, communicate the results, and explore correlation between automatic and human assessments.

Sampling

Human evaluation was carried out for eight submissions received during the WebNLG Challenge, for the baseline developed for the challenge, and for human references collected earlier while creating the WebNLG corpus [Gardent et al., 2017b]. Thus, we had ten systems in total including references, which we will refer to as WEBNLG. Each submission from a team included 1862 texts generated from data units. For human evaluation, we sampled 223 texts from each submission. A sample was chosen based on different characteristics of the WebNLG corpus: how many RDF triples were in data units (size from 1 to 5), and what was the DBpedia category (Building, City, Artist, etc.). To balance our sample, we also chose texts that had received different METEOR scores. The final sample for each team comprised texts from each category (15 texts); in each category all triple set sizes were covered (5 sizes), and finally for every category and every size, we extracted texts which got a low/medium/high sentence-level METEOR score when averaging scores across all teams. In such a way, our sample should have had 225 (i.e. $15 * 5 * 3$) texts; however, the count was reduced to 223, as one category (ComicsCharacter) had few data units for a particular size.

We calculated automatic evaluation metrics on our samples, and compared these rankings to those of the whole submissions (see Table 1). The last columns represent the initial rankings for all texts as reported in [Gardent et al., 2017a]. The METEOR ranking for a sample stayed the same as for all data; that could be expected as we sampled taking into account different METEOR scores. Rankings for BLEU and TER shifted a few teams along the ranking, but not drastically. Some of them gained or lost one point maximum.

Design

In total, we evaluated 2230 texts by collecting three judgments per text. Our participants came from English-speaking countries. They were shown data (set of RDF triples) and a system output (a text), and were asked to answer

Rank	Team	METEOR	Groups (on sample)	Groups (on all data)
1–2	UPF-FORGE	0.39	A	(A)
1–3	UTILBURG-SMT	0.38	A, B	(B)
2–3	UMELBOURNE	0.38	B	(C)
4–7	UTILBURG-NMT	0.34	C	(D)
4–7	ADAPT	0.34	C	(E)
4–7	PKUWRITER	0.33	C	(E)
4–7	UTILBURG-PIPELINE	0.32	C	(E)
8	BASELINE	0.26	D	(F)
9	UIT-VNU-HCM	0.08	E	(G)

Rank	Team	TER	Groups (on sample)	Groups (on all data)
1	UMELBOURNE	0.44	A	(A)
2–5	PKUWRITER	0.51	B	(C)
2–7	UTILBURG-SMT	0.52	B, C	(B)
2–7	UTILBURG-PIPELINE	0.53	B, C	(C)
2–7	UPF-FORGE	0.54	B, C	(C)
6–7	BASELINE	0.56	C	(D)
6–7	UTILBURG-NMT	0.57	C	(D)
8	ADAPT	0.72	D	(E)
9	UIT-VNU-HCM	0.84	E	(E)

Rank	Team	BLEU-4	Groups (on sample)	Groups (on all data)
1	UMELBOURNE	48.05	A	(A)
2–3	UTILBURG-SMT	45.90	B	(B)
2–3	PKUWRITER	43.71	B	(C)
4–6	UPF-FORGE	40.03	C	(C)
4–7	BASELINE	37.81	C, D	(E)
5–8	UTILBURG-PIPELINE	37.34	C, D	(D)
4–7	ADAPT	36.73	C, D	(F)
7–8	UTILBURG-NMT	35.98	E	(D)
9	UIT-VNU-HCM	5.25	F	(G)

Table 1: METEOR, TER, and BLEU rankings for sample data. The difference between systems which have a letter in common is not statistically significant ($\alpha = .05$). The last column denotes the initial team ranking for all data.

three questions:

- *Does the text correctly represent the meaning in the data?* (1 - Incorrectly, 2 - Medium, 3 - Correctly)
- *Rate the grammar and the spelling of the text: Is the text grammatical (no spelling or grammatical errors)?* (1 - Ungrammatical, 2 - Medium, 3 - Grammatical)
- *Rate the fluency of the text: Does the text sound fluent and natural?* (1 - Not fluent, 2 - Medium, 3 - Fluent)

The three questions with a three-point Lickert scale rate Semantic adequacy, Grammaticality, and Fluency respectively. One text with its corresponding data entry was shown per page. Each participant had a restriction to give only 30 answers per task. Texts were distributed by five separate tasks, which included outputs produced for the same size of data. We also ensured that each participant evaluated an equal number of texts per team, if possible.

Some rule-based system (UIT-VNU-HCM and UTILBURG-PIPELINE) outputs were empty for a particular data unit, so they were not presented for human evaluation. The lowest score "1" was attributed to those outputs for all assessed parameters.

Ensuring Quality

We use CrowdFlower[1] to collect human judgments. Apart from using obvious controlling techniques such as the time a contributor spends on a page, restricting a crowdworker to give a limited number of answers per task, we applied several checks to identify if there are untrustworthy workers ("spammers") or not. First, given a sufficient number of answers (say, more than ten), we eliminated contributors whose judgments have always the same pattern for all texts, for instance, "2-3-3" scores. Secondly, we made use of the MACE tool [Hovy et al., 2013] to identify unreliable crowdworkers. MACE allows to detect less trusted annotators in an unsupervised fashion by comparing the probability distributions of answers across annotators. Annotator reliability was calculated independently in three variables (Semantic adequacy, Grammaticality, and Fluency). Based on low ratings, we manually evaluated and eliminated spammers, and afterwards launched another round of collecting judgments to cover missing values. We did not trust MACE blindly, rather it was used as an indicator for examining a potentially bad

[1]https://www.crowdflower.com/

worker. There were cases when a participant demonstrated a high reliability while assessed in one variable (Fluency), whereas in another variable (Grammaticality) her reliability was low. Those participants were usually kept after examination. In such a way, we did not create a uniform distribution of answers, and tried to preserve a variety in human judgments.

Human Evaluation Results

Several judgments obtained for the same text were averaged, and final means for each evaluated variable are shown in Table 2. A Wilcoxon rank-sum test was carried out to establish a statistically significant difference between average scores of the systems[2]. We used a non-parametric test, since our sample data did not follow the normal distribution.

Global Scores. WEBNLG scored first for all tested variables. Having human references at the first place may serve as an indicator of the sanity of the human evaluation experiment. UPF-FORGE always follows WEBNLG, what allows us to say that its output is very close to human-produced sentences. Across three human ratings, teams are ranked more or less the same, except BASELINE and UTILBURG-SMT. BASELINE scores are lower for Grammar, while being in the middle for the other two ratings, whereas UTILBURG-SMT scores high in Grammar while showing moderate performance in Semantics and Fluency.

Scores on Seen and Unseen Categories. System ratings on seen and unseen categories are presented in Table 3 (more on that difference see in [Gardent et al., 2017a]). Showing the same trend as in the automatic evaluation results [Gardent et al., 2017a], ADAPT scores on seen categories are boosting, outperforming human references by a small margin in Semantics and Fluency. However, on unseen data, its performance is poor. As a general tendency, one can also notice that generation from unseen data gives predictably worse scores in all variables than generation from seen data.

Scores on Triple Set Sizes. Apart from exploring the ability to generate from out-of-domain data, the ability to generate human-like texts from different data sizes is also of interest. Ratings were calculated for each triple set size (from 1 to 5 triples). In Semantics and Fluency the decreasing trend across systems is present (see Figures 1, 3): the bigger the data size to generate from, the lower scores for Semantics and Fluency. However, it is also

[2]All statistical experiments in this report were conducted using R.

Semantics	Avg	Groups
WEBNLG	2.61	a
UPF-FORGE	2.47	b
MELBOURNE	2.39	c
PKUWRITER	2.39	c
BASELINE	2.36	c
ADAPT	2.31	c
TILB-PIPELINE	2.19	d
TILB-NMT	2.16	e
TILB-SMT	1.96	f
UIT-VNU	1.39	g

Grammar	Avg	Groups
WEBNLG	2.77	a
UPF-FORGE	2.68	b
TILB-SMT	2.42	c
ADAPT	2.30	c
MELBOURNE	2.30	d
TILB-PIPELINE	2.20	e
PKUWRITER	2.08	f
TILB-NMT	1.99	g
BASELINE	1.86	h
UIT-VNU	1.42	i

Fluency	Avg	Groups
WEBNLG	2.58	a
UPF-FORGE	2.34	b
PKUWRITER	2.34	b
MELBOURNE	2.27	cb
ADAPT	2.26	cb
BASELINE	2.25	c
TILB-PIPELINE	2.07	d
TILB-NMT	2.01	e
TILB-SMT	1.81	f
UIT-VNU	1.38	g

Table 2: Human evaluation average scores of Semantic adequacy, Grammaticality, and Fluency. Letters denote clusters of Wilcoxon rank-sum significance test ($\alpha = .05$). A colour for a team name indicates a type of the system used (NMT, SMT, Pipeline).

the case for WEBNLG which lets us suggest that Fluency for longer texts is more difficult to achieve when rendering multiple RDF triples to a text encompassing all of them. While the Grammar ranking stays the same for all data sizes for WEBNLG for most systems it drops as the size grows bigger. Nevertheless, UTILBURG-SMT and UTILBURG-PIPELINE demonstrate a small increase in their Grammar rating at bigger sizes.

Correlation between Evaluation Methods

To augment variability and allow comparisons with other evaluation studies, we perform correlation analysis both on system- and sentence-level. We use Spearman's correlation coefficient. For the sake of using multiple methods,

Semantics (seen)	Avg	Groups
ADAPT	2.66	a
WEBNLG	2.62	b
BASELINE	2.51	c
UPF-FORGE	2.50	c
PKUWRITER	2.50	c
MELBOURNE	2.46	d
TILB-PIPELINE	2.33	e
TILB-NMT	2.23	f
TILB-SMT	2.14	g
UIT-VNU	1.51	h

Semantics (unseen)	Avg	Groups
WEBNLG	2.59	a
UPF-FORGE	2.40	b
MELBOURNE	2.27	c
PKUWRITER	2.17	d
BASELINE	2.07	e
TILB-NMT	2.01	e
TILB-PIPELINE	1.91	f
ADAPT	1.63	g
TILB-SMT	1.61	g
UIT-VNU	1.14	h

Grammar (seen)	Avg	Groups
WEBNLG	2.76	a
ADAPT	2.73	a
UPF-FORGE	2.71	a
MELBOURNE	2.48	b
TILB-SMT	2.47	b
TILB-PIPELINE	2.46	b
PKUWRITER	2.23	c
BASELINE	2.21	c
TILB-NMT	2.14	c
UIT-VNU	1.55	d

Grammar (unseen)	Avg	Groups
WEBNLG	2.79	a
UPF-FORGE	2.62	a
TILB-SMT	2.30	b
MELBOURNE	1.93	c
PKUWRITER	1.77	c
TILB-PIPELINE	1.70	cd
TILB-NMT	1.69	cd
ADAPT	1.45	de
BASELINE	1.17	e
UIT-VNU	1.16	e

Fluency (seen)	Avg	Groups
ADAPT	2.62	a
WEBNLG	2.57	b
PKUWRITER	2.44	c
BASELINE	2.40	c
UPF-FORGE	2.36	d
MELBOURNE	2.35	d
TILB-PIPELINE	2.21	e
TILB-NMT	2.10	f
TILB-SMT	2.01	g
UIT-VNU	1.50	h

Fluency (unseen)	Avg	Groups
WEBNLG	2.61	a
UPF-FORGE	2.31	b
PKUWRITER	2.14	c
MELBOURNE	2.11	c
BASELINE	1.94	d
TILB-NMT	1.85	e
TILB-PIPELINE	1.78	f
ADAPT	1.55	g
TILB-SMT	1.44	h
UIT-VNU	1.14	i

Table 3: Human evaluation average scores for seen and unseen categories. Letters denote clusters of Wilcoxon rank-sum significance test ($\alpha = .05$). A colour for a team name indicates a type of the system used (NMT, SMT, Pipeline).

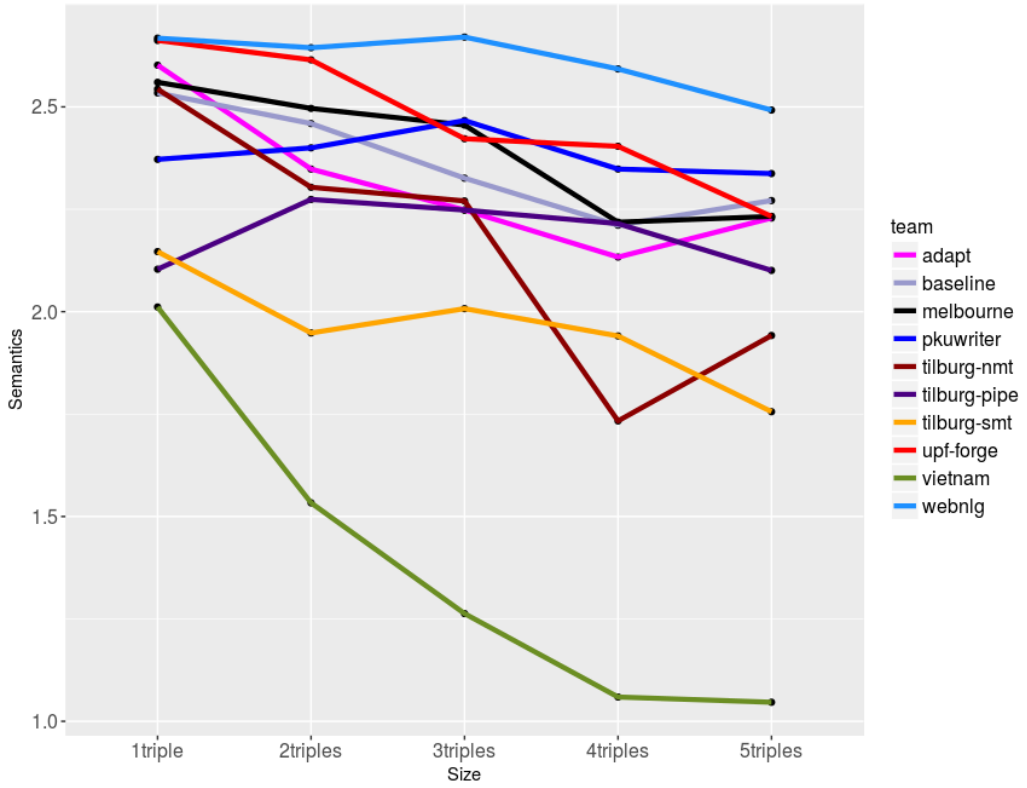

Figure 1: Semantics mean scores per size of data.

Pearson's correlation is reported in the Appendix. To prevent a possible bias, we excluded human references (WEBNLG) from the correlation analysis, since their automatic scores are equal to 1.0 (for BLEU and METEOR) and 0.0 (for TER). Thus, we have nine data points to build a regression line.

Automatic metrics were initially created to account for the evaluation of whole systems (i.e. they are corpus-based metrics). It is therefore unclear how applicable these are for the evaluation on a per-sentence basis. However, many studies reported correlation between human judgments and automatic metrics on sentence-level when they have one or few systems to evaluate (e.g. [Stent et al., 2005] for paraphrasing, [Elliott and Keller, 2014] for image caption generation, [Novikova et al., 2017] for NLG). In such a fashion, i.e. having more data points, statistical significance is easier to achieve. Since a research topic about the validity of automatic metrics versus human judgments is active and triggers a lot of discussions, we carried out a sentence-level correlation analysis as well. We hope it may be helpful to shed some light on the relationship between automatic and human judgments and to facilitate further comparing of numerous validation studies with different design.

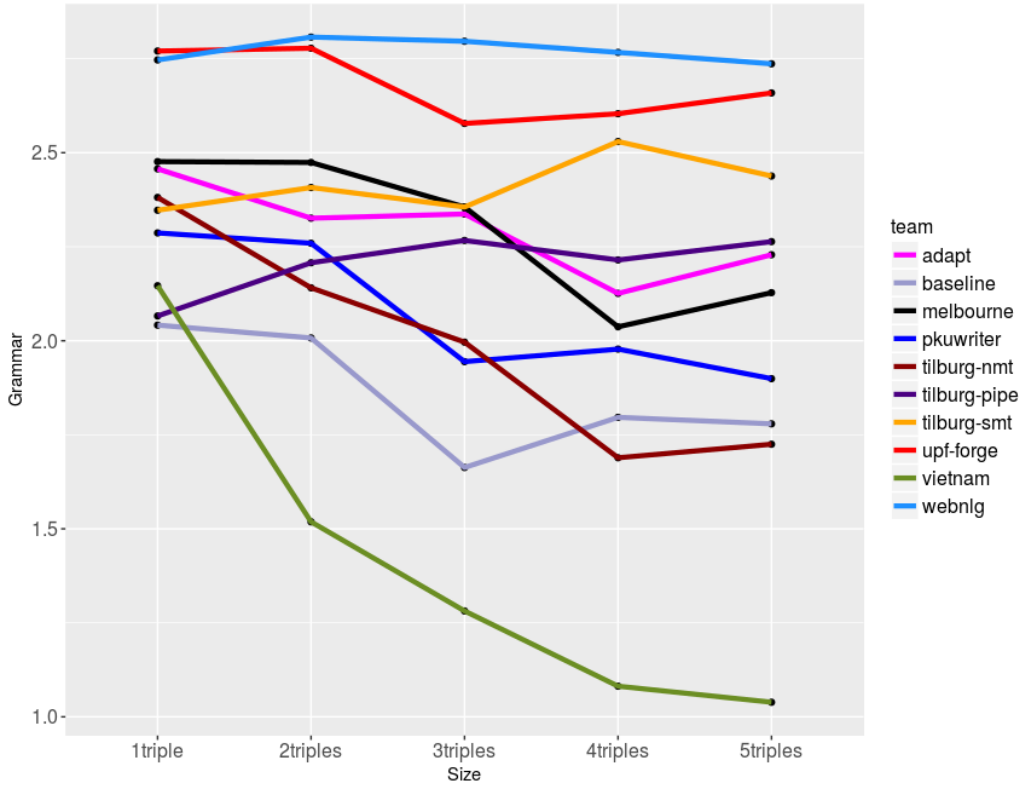

Figure 2: Grammar mean scores per size of data.

Figure 4 shows that statistically significant correlations ($p < .001$) were achieved only between semantics and METEOR if human vs automatic metrics relationship is of interest. METEOR exploration of stems and synonyms could well explain that strong correlation. Similar findings for METEOR were reported in the MT community [Callison-Burch et al., 2009] and in the image caption generation domain [Bernardi et al., 2016]. We also found strong correlation between TER and BLEU, and between judgments of grammar and fluency.

As for the sentence-level analysis (cf. Figure 5), all measurements demonstrated statistically significant correlations ($p < .001$). The highest correlation between human and automatic metrics was reached for METEOR and semantics ($\rho = 0.73$). For the rest of comparisons, correlations are moderate, ranging from $\rho = 0.49$ to $\rho = 0.59$. Also, automatic metrics show strong correlations with each other ($\rho \geq 0.78$).

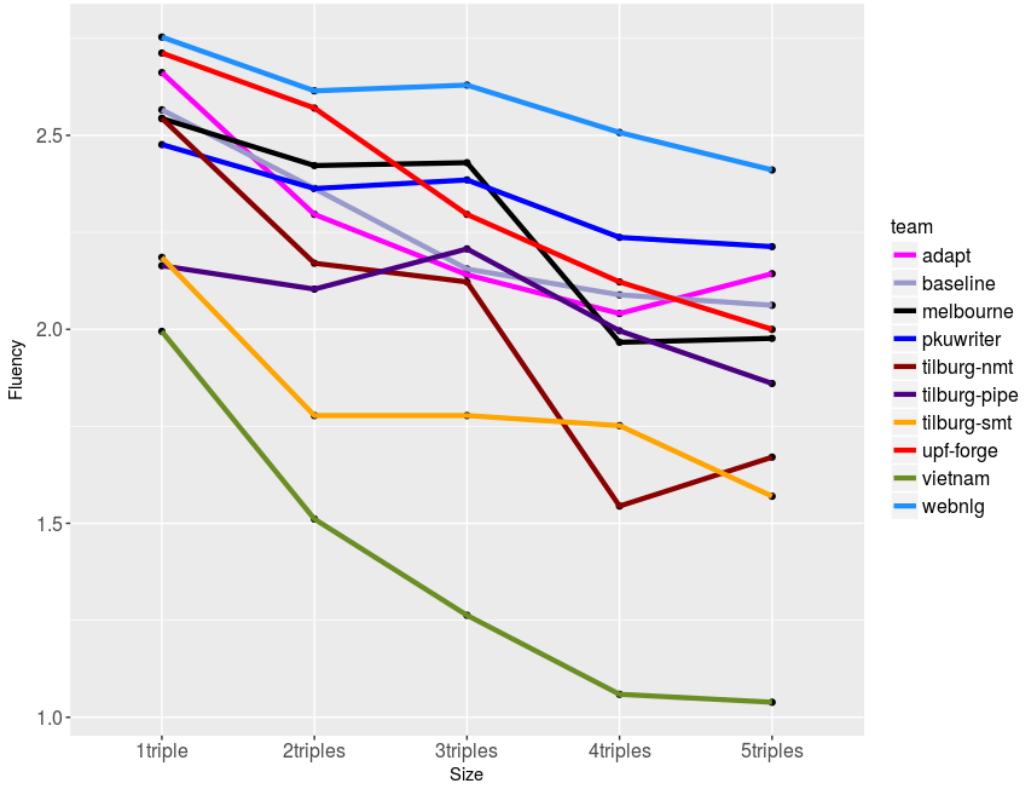

Figure 3: Fluency mean scores per size of data.

Conclusion

This report puts an end to the WebNLG Challenge, successfully ran between April and December 2017, announcing final human evaluation results. We paid a special attention to collect reliable human judgments. We carried out the test phase both on seen and unseen data, which allowed us to show that tuning a system on a given training data is not enough to develop a generic system, which will be able to perform equally on another type of semantic relations (RDF properties). We also underlined the importance of performing various types of correlation analysis between human and automatic metrics, showing that a design decision (using system- or sentence-level comparisons) influences a lot the outcome.

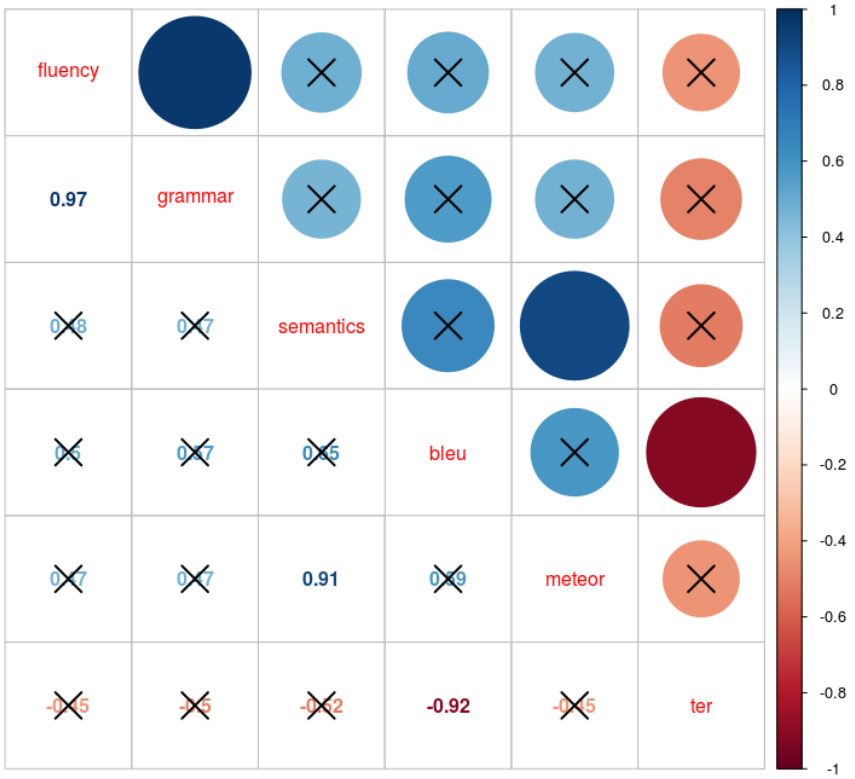

Figure 4: Spearman's correlation on system-level. Crossed squares indicate that statistical significance was not reached ($\alpha = .05$).

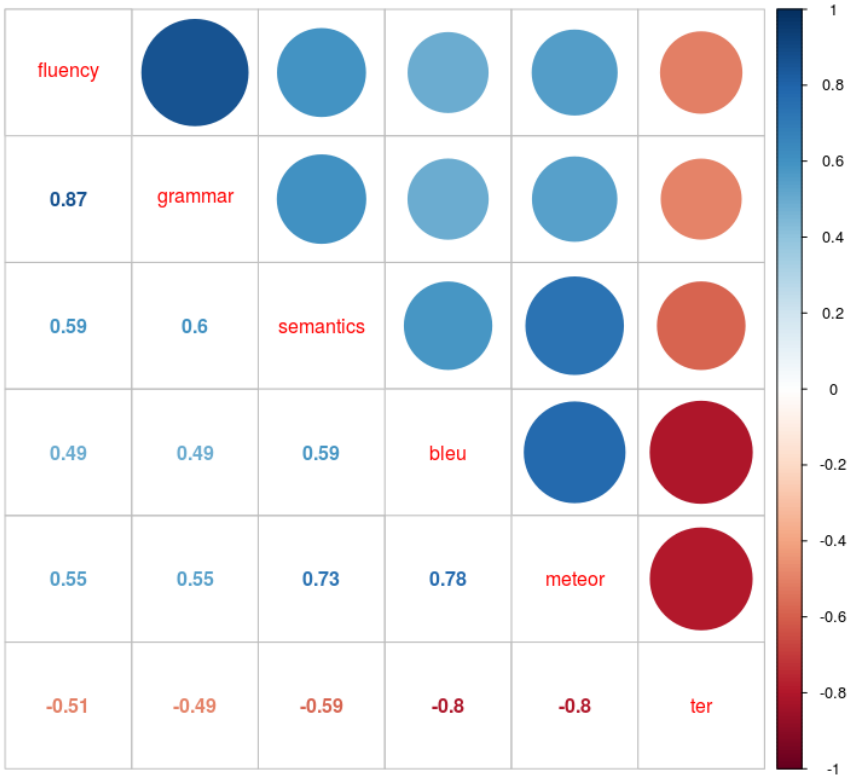

Figure 5: Spearman's correlation on sentence-level. All correlations are statistically significant ($\alpha = .001$).

Bibliography

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. The webnlg challenge: Generating text from rdf data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, 2017a.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, August 2017b. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H Hovy. Learning whom to trust with mace. In *HLT-NAACL*, pages 1120–1130, 2013.

Amanda Stent, Matthew Marge, and Mohit Singhai. Evaluating evaluation methods for generation in the presence of variation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 341–351. Springer, 2005.

Desmond Elliott and Frank Keller. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers*, volume 452, page 457, 2014.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, 2017.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*,

pages 1–28, Athens, Greece, March 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W09/W09-0401.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.(JAIR)*, 55:409–442, 2016.

Appendix

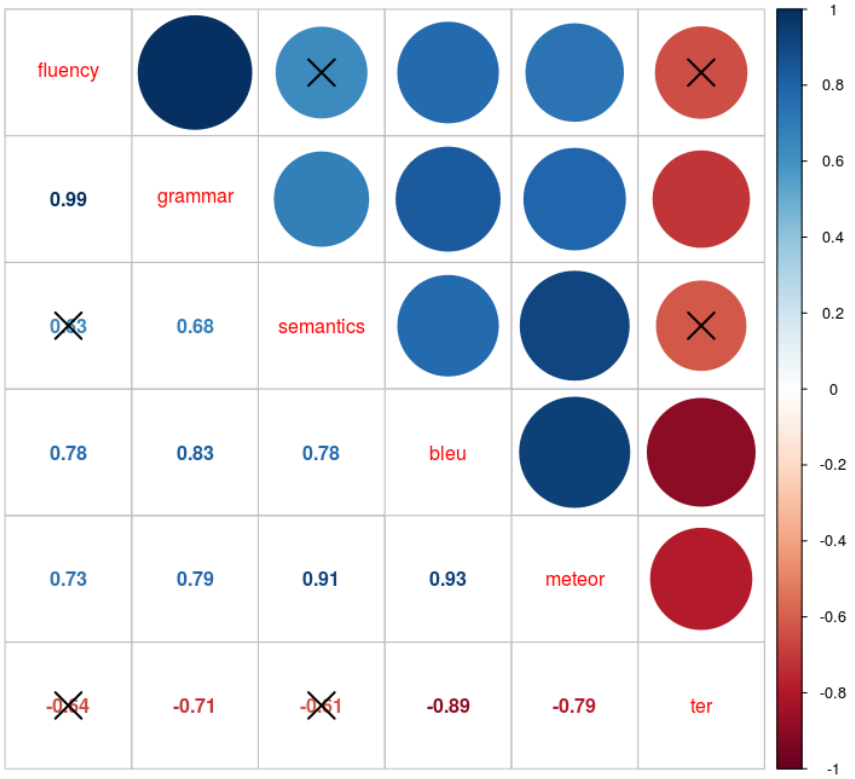

Figure 6: Pearson's correlation on system-level. Crossed squares indicate that statistical significance was not reached ($\alpha = .05$).

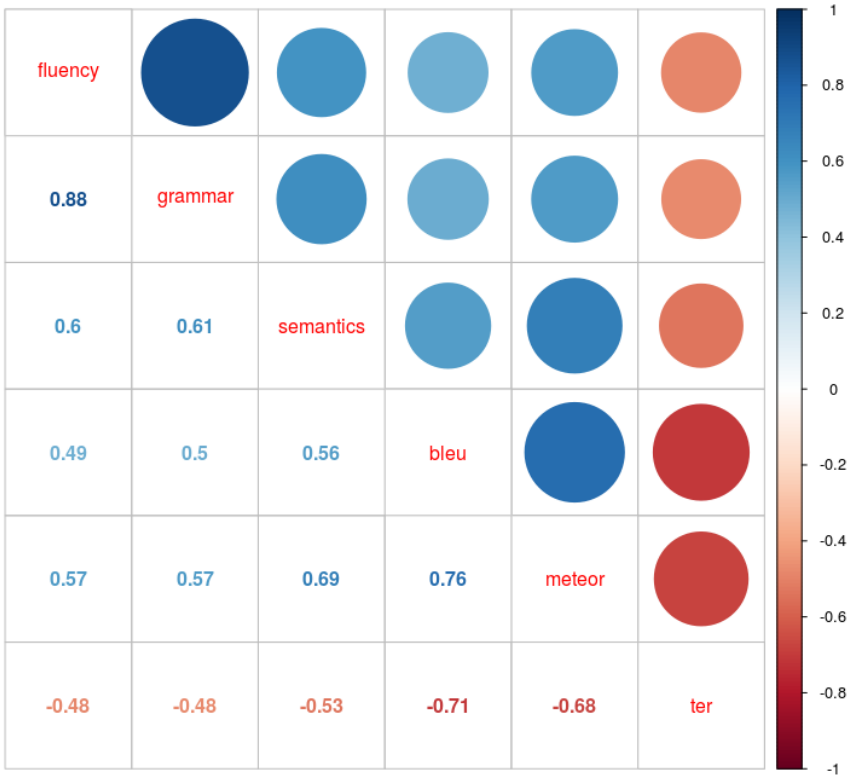

Figure 7: Pearson's correlation on sentence-level. All correlations are statistically significant ($\alpha = .001$).