Referenceless Quality Estimation for Natural Language Generation

Citation for published version:
Dusek, O, Novikova, J & Rieser, V 2017, Referenceless Quality Estimation for Natural Language Generation. in *Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.* ICML, 1st Workshop on Learning to Generate Natural Language, Sydney, Australia, 10/08/17.

Link:
Link to publication record in Heriot-Watt Research Portal

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017

Referenceless Quality Estimation for Natural Language Generation

Ondřej Dušek [1]   **Jekaterina Novikova** [1]   **Verena Rieser** [1]

Abstract

Traditional automatic evaluation measures for natural language generation (NLG) use costly human-authored references to estimate the quality of a system output. In this paper, we propose a referenceless quality estimation (QE) approach based on recurrent neural networks, which predicts a quality score for a NLG system output by comparing it to the source meaning representation only. Our method outperforms traditional metrics and a constant baseline in most respects; we also show that synthetic data helps to increase correlation results by 21% compared to the base system. Our results are comparable to results obtained in similar QE tasks despite the more challenging setting.

1. Introduction

Automatic evaluation of natural language generation (NLG) is a complex task due to multiple acceptable outcomes. Apart from manual human evaluation, most recent works in NLG are evaluated using word-overlap-based metrics such as BLEU (Gkatzia & Mahamood, 2015), which compute similarity against gold-standard human references. However, high quality human references are costly to obtain, and for most word-overlap metrics, a minimum of 4 references are needed in order to achieve reliable results (Finch et al., 2004). Furthermore, these metrics tend to perform poorly at segment level (Lavie & Agarwal, 2007; Chen & Cherry, 2014; Novikova et al., 2017a).

We present a novel approach to assessing NLG output quality without human references, focusing on segment-level (utterance-level) quality assessments.[1] We train a recurrent neural network (RNN) to estimate the quality of an

Figure 1. The architecture of our referenceless NLG QE model.

NLG output based on comparison with the source meaning representation (MR) only. This allows to efficiently assess NLG quality not only during system development, but also at runtime, e.g. for optimisation, reranking, or compensating low-quality output by rule-based fallback strategies.

To evaluate our method, we use crowdsourced human quality assessments of real system outputs from three different NLG systems on three datasets in two domains. We also show that adding fabricated data with synthesised errors to the training set increases relative performance by 21% (as measured by Pearson correlation).

In contrast to recent advances in referenceless quality estimation (QE) in other fields such as machine translation (MT) (Bojar et al., 2016) or grammatical error correction (Napoles et al., 2016), NLG QE is more challenging because (1) diverse realisations of a single MR are often acceptable (as the MR is typically a limited formal language); (2) human perception of NLG quality is highly variable, e.g. (Dethlefs et al., 2014); (3) NLG datasets are costly to obtain and thus small in size. Despite these difficulties, we achieve promising results – correlations with human judgements achieved by our system stay in a somewhat lower range than those achieved e.g. by state-of-the-art MT QE systems (Bojar et al., 2016), but they significantly outperform word-overlap metrics. To our knowledge, this is the first work in trainable NLG QE without references.

2. Our Model

We use a simple RNN model based on Gated Recurrent Units (GRUs) (Cho et al., 2014), composed of one GRU

---

[1]Interaction Lab, Heriot-Watt University, Edinburgh, Scotland, UK. Correspondence to: Ondřej Dušek <o.dusek@hw.ac.uk>.

[1]In our data, a "segment" refers to an utterance generated by an NLG system in the context of a human-computer dialogue, typically 1 or 2 sentences in length (see Section 3.1). We estimate the utterance quality without taking the dialogue context into ac-
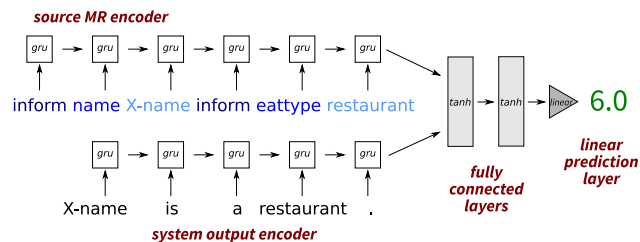
count. Assessing the appropriateness of responses in context is beyond the scope of this paper, see e.g. (Liu et al., 2016; Lowe et al., 2017; Cercas Curry et al., 2017).

encoder each for the source MR and the NLG system output to be rated, followed by fully connected layers operating over the last hidden states of both encoders. The final classification layer is linear and produces the quality assessment as a single floating point number (see Figure 1).[2]

The model assumes both the source MR and the system output to be sequences of tokens $x^M = \{x_1^M, \ldots, x_m^M\}$, $x^S = \{x_1^S, \ldots, x_n^S\}$, where each token is represented by its embedding (Bengio et al., 2003). The GRU encoders encode these sequences left-to-right into sequences of hidden states $\{h_t^M\}_{t=1}^m, \{h_t^S\}_{t=1}^n$:

$$h_x^M = \text{gru}(x_t^M, h_{t-1}^M) \quad (1)$$
$$h_x^S = \text{gru}(x_t^S, h_{t-1}^S) \quad (2)$$

The final hidden states h_m^M and h_n^S are then fed to a set of fully connected $\tanh$ layers $z_i, i \in \{0 \ldots k\}$:

$$z_0 = \tanh(W_0(h_m^M \circ h_n^S)) \quad (3)$$
$$z_i = \tanh(W_k z_{i-1}) \quad (4)$$

The final prediction is given by a linear layer:

$$\hat{y} = W_{k+1} z_k \quad (5)$$

In (3–5), $W_0 \ldots W_k$ stand for square weight matrices, W_{k+1} is a weight vector.

The network is trained in a supervised setting by minimising mean square error against human-assigned quality scores on the training data (see Section 3). Embedding vectors are initialised randomly and learned during training; each token found in the training data is given an embedding dictionary entry. Dropout (Hinton et al., 2012) is applied on the inputs to the GRU encoders for regularisation.

The floating point values predicted by the network are rounded to the precision and clipped to the range found in the training data.[3]

2.1. Model Variants

We also experimented with several variants of the model which performed similarly or worse than those presented in Section 4. We list them here for completeness:

- replacing GRU cells with LSTM (Hochreiter & Schmidhuber, 1997),
- using word embeddings pretrained by the word2vec tool (Mikolov et al., 2013) on Google News data,[4]

System↓ Data→	BAGEL	SFRest	SFHot	Total
LOLS	202	581	398	1,181
RNNLG	-	600	477	1,077
TGen	202	-	-	202
Total	404	1,181	875	2,460

Table 1. Number of ratings from different source datasets and NLG systems in our data.

- using a set of independent binary classifiers, each predicting one of the individual target quality levels (see Section 3.1),
- using an ordered set of binary classifiers trained to predict 1 for NLG outputs above a specified quality level, 0 below it,[5]
- pretraining the network using a different task (classifying MRs or predicting next word in the sentence).

3. Experimental Setup

In the following, we describe the data we use to evaluate our system, our method for data augmentation, detailed parameters of our model, and evaluation metrics.

3.1. Dataset

Using the CrowdFlower crowdsourcing platform,[6] we collected a dataset of human rankings for outputs of three recent data-driven NLG systems as provided to us by the systems' authors; see (Novikova et al., 2017a) for more details. The following systems are included in our set:

- LOLS (Lampouras & Vlachos, 2016), which is based on imitation learning,
- RNNLG (Wen et al., 2015), a RNN-based system,
- TGen (Dušek & Jurčíček, 2015), a system using perceptron-guided incremental tree generation.

Their outputs are on the test parts of the following datasets (see Table 1):

- BAGEL (Mairesse et al., 2010) – 404 short text segments (1–2 sentences) informing about restaurants,
- SFRest (Wen et al., 2015) – ca. 5,000 segments from the restaurant information domain (including questions, confirmations, greetings, etc.),
- SFHot (Wen et al., 2015) – a set from the hotel domain similar in size and contents to SFRest.

During the crowdsourcing evaluation of the outputs, crowd

	Instance	H	B	M	R	C	S4
MR	inform(name='la ciccia',area='bernal heights',price_range=moderate)	5.5	1	3	3.5	2	4.5
Ref	la ciccia is a moderate price restaurant in bernal heights		*0.000*	*0.371*	*0.542*	*2.117*	
Out	la ciccia, is in the bernal heights area with a moderate price range.						
MR	inform(name='intercontinental san francisco',price_range='pricey')	2	4.5	3	5.5	2	5
Ref	sure, the intercontinental san francisco is in the pricey range.		*0.707*	*0.433*	*0.875*	*2.318*	
Out	the intercontinental san francisco is in the pricey price range.						

Figure 2. Examples from our dataset. *Ref* = human reference (one selected), *Out* = system output to be rated. *H* = median human rating, *B* = BLEU, *M* = METEOR, *R* = ROUGE, *C* = CIDEr, *S4* = rating given by our system in the S4 configuration. The metrics are shown with normalised and rounded values (see Section 3.5) on top and the original, raw values underneath (in italics). The top example is rated low by all metrics, our system is more accurate. The bottom one is rated low by humans but high by some metrics and our system.

workers were given two random system outputs along with the source MRs and were asked to evaluate the absolute overall quality of both outputs on a 1–6 Likert scale (see Figure 2). We collected at least three ratings for each system output; this resulted in more ratings for the same sentence if two systems' outputs were identical. The ratings show a moderate inter-rater agreement of 0.45 ($p < 0.001$) intra-class correlation coefficient (Landis & Koch, 1977) across all three datasets. We computed medians of the three (or more) ratings in our experiments to ensure more consistent ratings, which resulted in .5 ratings for some examples. We keep this granularity throughout our experiments.

We use our data in a 5-fold cross-validation setting (three training, one development, and one test part in each fold). We also test our model on a subset of ratings for a particular NLG system or dataset in order to assess its cross-system and cross-domain performance (see Section 4).

3.2. Data Preprocessing

The source MRs in our data are variants of the dialogue acts (DA) formalism (Young et al., 2010) – a shallow domain-specific MR, consisting of the main DA type (*hello*, *inform*, *request*) and an optional set of slots (attributes, such as *food* or *location*) and values (e.g. *Chinese* for food or *city centre* for location). DAs are converted into sequences for our system as a list of triplets "DA type – slot – value", where DA type may be repeated multiple times and/or special null tokens are used if slots or values are not present (see Figure 1). The system outputs are tokenised and lowercased for our purposes.

We use delexicalisation to prevent data sparsity, following (Mairesse et al., 2010; Henderson et al., 2014; Wen et al., 2015), where values of most DA slots (except unknown and binary *yes/no* values) are replaced in both the source MRs and the system outputs by slot placeholders – e.g. *Chinese* is replaced by *X-food* (cf. also Figure 1).[7]

---

[7]Note that only values matching the source MR are delexicalised in the system outputs – if the outputs contain values not present in the source MR, they are kept intact in the model input.

Setup	Instances
S1	1,476
S2	3,937
S3	13,442
S4	45,137
S5	57,372
S6	80,522

Table 2. Training data size comparison for different data augmentation procedures (varies slightly in different cross-validation folds due to rounding).

3.3. Synthesising Additional Training Data

Following prior works in grammatical error correction (Rozovskaya et al., 2012; Felice & Yuan, 2014; Xie et al., 2016), we synthesise additional training instances by introducing artificial errors: Given a training instance (source MR, system output, and human rating), we generate a number of errors in the system output and lower the human rating accordingly. We use a set of basic heuristics mimicking some of the observed system behaviour to introduce errors into the system outputs:[8]

1. removing a word,

2. duplicating a word at its current position,

3. duplicating a word at a random position,

4. adding a random word from a dictionary learned from all training system outputs,

5. replacing a word with a random word from the dictionary.

We lower the original Likert scale rating of the instance by 1 for each generated error. If the original rating was 5.5 or 6, the rating is lowered to 4 with the first introduced error and by 1 for each additional error.

---

[8]To ensure that the errors truly are detrimental to the system output quality, our rules prioritise content words, i.e. they do not change articles or punctuation if other words are present. The rules never remove the last word left in the system output.

We also experiment with using additional natural language sentences where human ratings are not available – we use human-authored references from the original training datasets and assume that these would receive the maximum rating of 6. We introduce artificial errors into the human references in exactly the same way as with training system outputs.

3.4. Model Training Parameters

We set the network parameters based on several experiments performed on the development set of one of the cross-validation folds (see Section 3.1).[9]

We train the network for 500 passes over the training data, checking Pearson and Spearman correlations on the validation set after each pass (with equal importance). We keep the configuration that yielded the best correlations overall. For setups using synthetic training data (see Section 3.3), we first perform 20 passes over all data including synthetic, keeping the best parameters, and then proceed with 500 passes over the original data. To compensate for the effects of random network initialisation, all our results are averaged over five runs with different initial random seeds following Wen et al. (2015).

3.5. Evaluation Measures

Following practices from MT Quality estimation (Bojar et al., 2016),[10] we use Pearson's correlation of the predicted output quality with median human ratings as our primary evaluation metric. Mean absolute error (MAE), root mean squared error (RMSE), and Spearman's rank correlation are used as additional metrics.

We compare our results to some of the common word-overlap metrics – BLEU (Papineni et al., 2002), METEOR (Lavie & Agarwal, 2007), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015) – normalised into the 1–6 range of the predicted human ratings and further rounded to 0.5 steps.[11] In addition, we also show the MAE/RMSE values for a trivial constant baseline that always predicts the overall average human rating (4.5).

4. Results

We test the following configurations that only differ in the amount and nature of synthetic data used (see Section 3.3

and Table 2):

S1 Base system variant, with no synthetic data.

S2 Adding synthetic data – introducing artificial errors into system outputs from the training portion of our dataset (no additional human references are used).

S3 Same as previous, but with additional human references from the training portion of our dataset (including artificial errors; see Section 3.3).[12]

S4 As previous, but with additional human references from the training parts of the respective source NLG datasets (including artificial errors), i.e. references on which the original NLG systems were trained.[12]

S5 As previous, but also includes additional human references from the test portion of our dataset (including artificial errors).[13]

S6 As previous, but also includes development parts of the source NLG datasets (including artificial errors).

Synthetic data are never created from system outputs in the test part of our dataset. Note that S1 and S2 only use the original system outputs and their ratings, with no additional human references. S3 and S4 use additional human references (i.e. more in-domain data), but do not use human references for the instances on which the system is tested. S5 and S6 also use human references for test MRs, even if not directly, and are thus not strictly referenceless.

4.1. Results using the whole dataset

The correlations and error values we obtained over the whole data in a cross-validation setup are shown in Table 3. The correlations only stay moderate for all system variants. On the other hand, we can see that even the base setup (S1) trained using less than 2,000 examples performs better than all the word-overlap metrics in terms of all evaluation measures. Improvements in both Pearson and Spearman correlations are significant according to the Williams test (Williams, 1959; Kilickaya et al., 2017) ($p < 0.01$). When comparing the base setup against the constant baseline, MAE is lower but RMSE is slightly higher, which suggests that our system does better on average but is prone to occasional large errors.

The results also show that the performance can be improved considerably by adding synthetic data, especially after more than tripling the training data in S4 (Pearson correlation improvements are statistically significant in terms

---

[9]We use embedding size 300, learning rate 0.0001, dropout probability 0.5, and two fully connected $\tanh$ layers ($k = 2$).

[10]See also the currently ongoing WMT'17 Quality Estimation shared task at http://www.statmt.org/wmt17/quality-estimation-task.html.

[11]We used the Microsoft COCO Captions evaluation script to obtain the metrics scores (Chen et al., 2015). Trials with non-quantised metrics yielded very similar correlations.

---

[12]As mentioned in Section 3.1, our dataset only comprises the test sets of the source NLG datasets, i.e. the additional human references in S3 represent a portion of the source test sets. The difference to S4 is the amount of the additional data (see Table 2).

[13]Note that the model still does not have any access at training time to test NLG system outputs or their true ratings.

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: + errors generated in training system outputs	0.283	0.268	0.948	1.273
S3: + training references, with generated errors	0.278	0.261	0.930	1.257
S4: + systems training data, with generated errors	**0.330**	0.274	0.914	1.226
S5: + test references, with generated errors*	**0.331**	0.265	0.937	1.245
S6: + complete datasets, with generated errors*	**0.354**	**0.287**	0.909	1.208

Table 3. Results using cross-validation over the whole dataset. Setups marked with "*" use human references for the test instances. All setups S1–S6 produce significantly better correlations than all metrics ($p < 0.01$). Significant improvements in correlation ($p < 0.05$) over S1 are marked in bold.

of the Williams test, $p < 0.01$). Using additional human references for the test data seems to be helping further in S6 (the difference in Pearson correlation is statistically significant, $p < 0.01$): The additional references apparently provide more information even though the SFHot and SFRest datasets have similar MRs (identical when delexicalised) in training and test data (Lampouras & Vlachos, 2016).[14] The setups using larger synthetic data further improve MAE and RMSE: S4 and S6 increase the margin against the constant baseline up to ca. 0.1 in terms of MAE, and both are able to surpass the constant baseline in terms of RMSE.

4.2. Cross-domain and Cross-System Training

Next, we test how well our approach generalises to new systems and datasets and how much in-set data (same domain/system) is needed to obtain reasonable results. We use the SFHot data as our test domain and LOLS as our test system, and we treat the rest as out-of-set. We test three different configurations:

C1 Training exclusively using a small amount of in-set data (200 instances, 100 reserved for validation), testing on the rest of the in-set.

C2 Training and validating exclusively on out-of-set data, testing on the same part of the in-set as in C1 and C3.

C3 Training on the out-of-set data with a small amount of in-set data (200 instances, 100 reserved for validation), testing on the rest of the in-set.

---

[14]Note that unlike in NLG systems training on SFHot and SFRest, the identical MRs in training and test data do not allow our system to only memorize the training data as the NLG outputs to be rated are distinct. However, the situation is not 100% referenceless as the system may have been exposed to other NLG outputs for the same MR. Our preliminary experiments suggest that our systems can also handle lexicalised data well, without any modification (Pearson correlation 0.264–0.359 for S1–S6).

The results are shown in Tables 4 and 5, respectively.

The correlations of C2 suggest that while our network can generalise across systems to some extent (if data fabrication is used), it does not generalise well across domains without using in-domain training data. C1 and C3 configuration results demonstrate that even small amounts of in-set data help noticeably. However, if in-set data is used, additional out-of-set data does not improve the results in most cases (C3 is mostly not significantly better than the corresponding C1).

Except for a few cross-system C2 configurations with low amounts of synthetic data, all systems perform better than word-overlap metrics. However, most setups are not able to improve over the constant baseline in terms of MAE and RMSE.

5. Related Work

This work is the first NLG QE system to our knowledge; the most related work in NLG is probably the system by Dethlefs et al. (2014), which reranks NLG outputs by estimating their properties (such as colloquialism or politeness) using various regression models. However, our work is also related to QE research in other areas, such as MT (Specia et al., 2010), dialogue systems (Lowe et al., 2017) or grammatical error correction (Napoles et al., 2016). QE is especially well researched for MT, where regular QE shared tasks are organised (Callison-Burch et al., 2012; Bojar et al., 2013; 2014; 2015; 2016).

Many of the past MT QE systems participating in the shared tasks are based on Support Vector Regression (Specia et al., 2015; Bojar et al., 2014; 2015). Only in the past year, NN-based solutions started to emerge. Patel & M (2016) present a system based on RNN language models, which focuses on predicting MT quality on the word level.

	C1: small in-domain data only				C2: out-of-domain data only				C3: out-of-dom. + small in-dom.			
	Pear	**Spea**	**MAE**	**RMSE**	**Pear**	**Spea**	**MAE**	**RMSE**	**Pear**	**Spea**	**MAE**	**RMSE**
Constant	-	-	0.994	1.224	-	-	0.994	1.224	-	-	0.994	1.224
BLEU*	0.033	0.016	2.235	2.710	0.033	0.016	2.235	2.710	0.033	0.016	2.235	2.710
METEOR*	0.076	0.074	1.719	2.034	0.076	0.074	1.719	2.034	0.076	0.074	1.719	2.034
ROUGE-L*	0.064	0.049	1.255	1.620	0.064	0.049	1.255	1.620	0.064	0.049	1.255	1.620
CIDEr*	0.048	0.043	2.590	2.921	0.048	0.043	2.590	2.921	0.048	0.043	2.590	2.921
S1	0.147	0.136	1.086	1.416	0.162	0.152	0.985	1.281	0.170	<u>0.166</u>	1.003	1.315
S2	**0.196**	**0.176**	1.059	1.364	**0.197**	**0.189**	1.003	1.311	**0.219**	<u>**0.218**</u>	0.988	1.296
S3	**0.176**	**0.163**	1.093	1.420	0.147	0.134	1.037	1.366	<u>**0.219**</u>	<u>**0.216**</u>	0.979	1.302
S4	**0.264**	**0.218**	0.983	1.307	0.162	0.138	1.084	1.448	**0.247**	**0.211**	0.983	1.306
S5*	**0.280**	**0.221**	1.009	1.341	0.173	0.145	1.077	1.438	**0.210**	0.162	1.095	1.442
S6*	**0.271**	**0.202**	0.991	1.331	0.188	0.178	1.037	1.392	**0.224**	**0.210**	1.002	1.339

Table 4. Cross-domain evaluation results. Setups marked with "*" use human references of test instances. All setups produce significantly better correlations than all metrics ($p < 0.01$). Significant improvements in correlation ($p < 0.05$) over the corresponding S1 are marked in bold, significant improvements over the corresponding C1 are underlined.

	C1: small in-system data only				C2: out-of-system data only				C3: out-of-sys. + small in-sys.			
	Pear	**Spea**	**MAE**	**RMSE**	**Pear**	**Spea**	**MAE**	**RMSE**	**Pear**	**Spea**	**MAE**	**RMSE**
Constant	-	-	1.060	1.301	-	-	1.060	1.301	-	-	1.060	1.301
BLEU*	0.079	0.043	2.514	2.971	0.079	0.043	2.514	2.971	0.079	0.043	2.514	2.971
METEOR*	0.141	0.122	1.929	2.238	0.141	0.122	1.929	2.238	0.141	0.122	1.929	2.238
ROUGE-L*	0.064	0.048	1.449	1.802	0.064	0.048	1.449	1.802	0.064	0.048	1.449	1.802
CIDEr*	0.127	0.106	2.801	3.112	0.127	0.106	2.801	3.112	0.127	0.106	2.801	3.112
S1	*0.341*	0.334	1.054	1.405	*0.097*	*0.117*	1.052	1.336	0.174	0.179	1.114	1.455
S2	**0.358**	0.345	1.007	1.342	*0.115*	*0.119*	1.057	1.355	**0.203**	**0.222**	1.253	1.613
S3	**0.378**	**0.365**	0.971	1.326	*0.112*	*0.094*	1.059	1.387	<u>**0.404**</u>	**0.377**	0.968	1.277
S4	**0.390**	**0.360**	0.981	1.311	**0.247**	**0.189**	1.011	1.338	**0.370**	**0.346**	0.997	1.312
S5*	**0.398**	**0.364**	1.043	1.393	**0.229**	**0.174**	1.025	1.328	**0.386**	**0.356**	0.975	1.301
S6*	**0.390**	0.353	1.036	1.389	**0.332**	**0.262**	0.969	1.280	**0.374**	**0.330**	0.979	1.298

Table 5. Cross-system evaluation results. Setups marked with "*" use human references of test instances. Setups that do not produce significantly better correlations than all metrics ($p < 0.05$) are marked in italics. Significant improvements in correlation ($p < 0.05$) over the corresponding S1 are marked in bold, significant improvements over the corresponding C1 are underlined.

Kim & Lee (2016) estimate segment-level MT output quality using a bidirectional RNN over both source and output sentences combined with a logistic prediction layer. They pretrain their RNN on large MT training data.

Last year's MT QE shared task systems achieve Pearson correlations between 0.4–0.5, which is slightly higher than our best results. However, the results are not directly comparable: First, we predict a Likert-scale assessment instead of the number of required post-edits. Second, NLG datasets are considerably smaller than corpora available for MT. Third, we believe that QE for NLG is harder due to the reasons outlined in Section 1.

6. Conclusions and Future Work

We presented the first system for referenceless quality estimation of natural language generation outputs. All code and data used here is available online at:

https://github.com/tuetschek/ratpred

In an evaluation spanning outputs of three different NLG systems and three datasets, our system significantly outperformed four commonly used reference-based metrics. It also improved over a constant baseline, which always predicts the mean human rating, in terms of MAE and RMSE. The smaller RMSE improvement suggests that our system is prone to occasional large errors. We have shown that generating additional training data, e.g. by using NLG training datasets and synthetic errors, significantly improves the system performance. While our system can generalise to unseen NLG systems in the same domain to some extent, its cross-domain generalisation capability is poor. However, very small amounts of in-domain/in-system data improve performance notably.

In future work, we will explore improvements to our error synthesising methods as well as changes to our network architecture (bidirectional RNNs or convolutional NNs). We also plan to focus on relative ranking of different NLG outputs for the same source MR or predicting the number of post-edits required. We intend to use data collected within the ongoing E2E NLG Challenge (Novikova et al., 2017b), which promises greater diversity than current datasets.

Acknowledgements

References

Bengio, Yoshua, Ducharme, Réjean, Vincent, Pascal, and Jauvin, Christian. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155, February 2003.

Bojar, Ondrej, Chatterjee, Rajen, Federmann, Christian, Graham, Yvette, Haddow, Barry, Huck, Matthias, Yepes, Antonio Jimeno, Koehn, Philipp, Logacheva, Varvara, Monz, Christof, and others. Findings of the 2016 conference on machine translation (WMT16). In *Proceedings of the First Conference on Machine Translation (WMT)*, pp. 131–198, Berlin, Germany, 2016.

Bojar, Ondřej, Buck, Christian, Callison-Burch, Chris, Federmann, Christian, Haddow, Barry, Koehn, Philipp, Monz, Christof, Post, Matt, Soricut, Radu, and Specia, Lucia. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pp. 1–44, Sofia, Bulgaria, 2013.

Bojar, Ondřej, Buck, Christian, Federmann, Christian, Haddow, Barry, Koehn, Philipp, Leveling, Johannes, Monz, Christof, Pecina, Pavel, Post, Matt, Saint-Amand, Herve, Soricut, Radu, Specia, Lucia, and Tamchyna, Aleš. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, Baltimore, MD, USA, 2014.

Bojar, Ondřej, Chatterjee, Rajen, Federmann, Christian, Haddow, Barry, Huck, Matthias, Hokamp, Chris, Koehn, Philipp, Logacheva, Varvara, Monz, Christof, Negri, Matteo, Post, Matt, Scarton, Carolina, Specia, Lucia, and Turchi, Marco. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 1–46, Lisbon, Portugal, 2015.

Callison-Burch, Chris, Koehn, Philipp, Monz, Christof, Post, Matt, Soricut, Radu, and Specia, Lucia. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 10–51, Montréal, Canada, 2012.

Cercas Curry, Amanda, Hastie, Helen, and Rieser, Verena. A review of evaluation techniques for social dialogue systems. Under submission, 2017.

Chen, Boxing and Cherry, Colin. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 362–367, Baltimore, MD, USA, 2014.

Chen, Xinlei, Fang, Hao, Lin, Tsung-Yi, Vedantam, Ramakrishna, Gupta, Saurabh, Dollar, Piotr, and Zitnick, C. Lawrence. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv:1504.00325 [cs]*, April 2015.

Cho, Kyunghyun, van Merrienboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, 2014. arXiv: 1406.1078.

Dethlefs, Nina, Cuayáhuitl, Heriberto, Hastie, Helen F, Rieser, Verena, and Lemon, Oliver. Cluster-based prediction of user ratings for stylistic surface realisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 702–711, Gothenburg, Sweden, 2014.

Dušek, Ondřej and Jurčíček, Filip. Training a Natural Language Generator From Unaligned Data. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 451–461, Beijing, China, 2015.

Felice, Mariano and Yuan, Zheng. Generating artificial errors for grammatical error correction. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 116–126, Gothenburg, Sweden, 2014.

Finch, Andrew M, Akiba, Yasuhiro, and Sumita, Eiichiro. How does automatic machine translation evaluation correlate with human scoring as the number of reference translations increases? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pp. 2019–2022, Lisbon, Portugal, 2004.

Gkatzia, Dimitra and Mahamood, Saad. A Snapshot of NLG Evaluation Practices 2005 - 2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pp. 57–60, Brighton, UK, 2015.

Henderson, M., Thomson, B., and Young, S. Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. In *Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT)*, pp. 360–365, South Lake Tahoe, NV, USA, 2014.

Hinton, Geoffrey E., Srivastava, Nitish, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580 [cs]*, July 2012.

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Kilickaya, Mert, Erdem, Aykut, Ikizler-Cinbis, Nazli, and Erdem, Erkut. Re-evaluating Automatic Metrics for Image Captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 199–209, Valencia, Spain, 2017. arXiv:1612.07600.

Kim, Hyun and Lee, Jong-Hyeok. Recurrent Neural Network based Translation Quality Estimation. In *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*, pp. 787–792, Berlin, Germany, 2016.

Lampouras, Gerasimos and Vlachos, Andreas. Imitation learning for language generation from unaligned data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1101–1112, Osaka, Japan, 2016.

Landis, J Richard and Koch, Gary G. The measurement of observer agreement for categorical data. *Biometrics*, 33 (1):159–174, March 1977.

Lavie, Alon and Agarwal, Abhaya. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 228–231, Prague, Czech Republic, 2007.

Lin, Chin-Yew. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pp. 74–81. Barcelona, Spain, 2004.

Liu, Chia-Wei, Lowe, Ryan, Serban, Iulian, Noseworthy, Michael, Charlin, Laurent, and Pineau, Joelle. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2122–2132, Austin, TX, USA, 2016. arXiv:1603.08023.

Lowe, Ryan, Noseworthy, Michael, Serban, Iulian V., Angelard-Gontier, Nicolas, Bengio, Yoshua, and Pineau, Joelle. Towards an automatic Turing test: learning to evaluate dialogue responses. In *Proceedings of 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1116–1126, Vancouver, Canada, 2017.

Mairesse, F., Gašić, M., Jurčíček, F., Keizer, S., Thomson, B., Yu, K., and Young, S. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1552–1561, Uppsala, Sweden, 2010.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv:1301.3781 [cs]*, January 2013.

Napoles, Courtney, Sakaguchi, Keisuke, and Tetreault, Joel. There's No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2109–2115, Austin, TX, USA, 2016.

Novikova, Jekaterina, Cercas Curry, Amanda, Dušek, Ondřej, and Rieser, Verena. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017a. arXiv:1707.06875.

Novikova, Jekaterina, Dušek, Ondřej, and Rieser, Verena. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Saarbrücken, Germany, 2017b. arXiv:1706.09254.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, Pennsylvania, PA, USA, 2002.

Patel, Raj Nath and M, Sasikumar. Translation Quality Estimation using Recurrent Neural Network. In *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*, pp. 819–824, Berlin, Germany, 2016.

Rozovskaya, Alla, Sammons, Mark, and Roth, Dan. The UI System in the HOO 2012 Shared Task on Error Correction. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pp. 272–280, Montréal, Canada, 2012.

Specia, Lucia, Raj, Dhwaj, and Turchi, Marco. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50, March 2010.

Specia, Lucia, Paetzold, Gustavo, and Scarton, Carolina. Multi-level Translation Quality Prediction with QuEst++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pp. 115–120, Beijing, China, 2015.

Vedantam, Ramakrishna, Lawrence Zitnick, C., and Parikh, Devi. CIDEr: Consensus-Based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, Boston, MA, USA, 2015. arXiv:1411.5726.

Wen, Tsung-Hsien, Gasic, Milica, Mrkšić, Nikola, Su, Pei-Hao, Vandyke, David, and Young, Steve. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1711–1721, Lisbon, Portugal, 2015.

Williams, Evan James. *Regression analysis*. Wiley, New York, NY, USA, 1959.

Xie, Ziang, Avati, Anand, Arivazhagan, Naveen, Jurafsky, Dan, and Ng, Andrew Y. Neural Language Correction with Character-Based Attention. *arXiv:1603.09727 [cs]*, March 2016.

Young, Steve, Gašić, Milica, Keizer, Simon, Mairesse, François, Schatzmann, Jost, Thomson, Blaise, and Yu, Kai. The Hidden Information State model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174, April 2010.