Affective Natural Language Generation by Phrase Insertion

Tomasz Dryjański*, Paweł Bujnowski*, Hyungtak Choi†, Katarzyna Podlaska*,
Kamil Michalski*, Katarzyna Beksa* and Paweł Kubik
*Samsung R&D Institute Poland
†Samsung Electronics, Korea
Emails: t.dryjanski, p.bujnowski, ht777.choi, k.podlaska,
k.michalski2, k.beksa @samsung.com; p.kubik.pl@gmail.com

***Abstract*—We propose a highly precise, production-ready neural model for affective natural language generation. It is designed to add predefined sentiment to neutral utterances without changing the meaning significantly. It works by inferring phrases and their insertion points. In our work we also propose strict correctness criteria and apply them to our inference results achieving human-level precision.**

The model is not specific to any particular domain like *IoT* or *restaurants review*. We use six selected emotion categories, but we also speculate that the model could be applied to other affective categories, like *informal style* or *politeness*, without a design change.

1. INTRODUCTION

Recently we observe a growing interest in the AI community towards emotions in the natural language processing. This holds true for affective natural language generation (NLG) in particular. Gatt and Krahmer [1] have lately compiled a survey on NLG. They roughly distinguish two groups in the existing solutions: rule-based, and applying some sort of machine learning, including deep learning (DL). Certainly other systems also exist, including hybrids. To the best of our knowledge, state-of-the-art DL models applied to NLG still lack precision, and they unacceptably often produce non-grammatical or nonsense utterances. For this reason they cannot be directly used in production systems. On the other hand rule-based systems require significant effort during creation, and they are difficult to maintain.

In our paper we discuss the task of applying a predefined emotional state of a speaker to a given neutral sentence. We propose a text-to-text neural model capable of emotional phrases insertion without changing the utterance meaning. The model is intended to be safe to use and to avoid handcrafted rules. Moreover, our choice of emotions for the model is somehow arbitrary, meaning that the set can be extended or fully replaced with no design change. It may be adapted to other NLG tasks that can be defined in terms of multipoint modifications, i.e. not only affective. It is designed to be fully general, so it is not bound to any specific domain like *IoT* or *restaurants review*.

The paper is organized as follows: we review the existing research in Section 2, and present our solution at the conceptual level in Section 3. Then we discuss the neural model in detail in Section 4, and proceed to the experimental protocol in Section 5. We present the results in Section 6 and conclude our work in Section 7.

2. RELATED WORK

The Affective Natural Language Generation (ANLG) term was originally coined by De Rosis & Grasso [2]. They used the rule-based approach that dominated the NLG field at that time [3][4]. Also pattern-based models [5][6] were used for this purpose. However, in the last few years, ANLG has witnessed renewed interest among researchers working on neural approaches to generation [1]. This has been motivated by the fact that rule-based and similar systems are hard to build and maintain.

Nowadays we observe a wide use of the sequence-to-sequence (Seq2Seq) framework [7] in text-to-text NLG tasks. Ashgar et al. [8] use it for affective neural dialogue generation. From their results we cannot directly determine how many utterances were actually rated as incorrect, which is the key theme of our work. Their task is also different from ours in that it does not require keeping the original utterance meaning unchanged. Other neural approaches to NLG–including those not related to affect–admit the possibility of incorrect inferences and rather focus on averaged results [9][10]. Ghosh et al. [11] regarded the affect strength as a conditioning factor similarly to our approach. The difference is that they trained their model for the task of completing utterances conditioned on their beginnings.

The abovementioned models belong to the generalized Seq2Seq family, a method susceptible to the *exposure bias* [12]. It occurs in a sequenced generation when a model is only exposed to the training data distribution instead of its own predictions. We avoid this issue by calculating all phrase insertions as a single inference step, i.e. not conditioning them one on another. Doing so, we also significantly reduce the unnecessary degrees of freedom. The approach was originally inspired by Pointer Networks [13] and their capability to infer an index from a variable-sized sequence.

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

Initially we focused on single-point phrase insertions, but ultimately we extended the model to multipoint for the sake of flexibility and naturalness [14].

To the best of our knowledge no publication proposes a generative model meeting the aforementioned strict correctness criteria and keeping the utterance meaning. This also holds true for NLG tasks not dealing with sentiment.

3. HIGH LEVEL SOLUTION OVERVIEW

The approach based on phrase insertions is not as simplistic as it might look at the first glance. The following examples

Do you think we can **still** be friends**, you fool**?
Wow! Do you think we can be friends?
Do you think we can be **fucking** friends?

illustrate that the phrase location cannot be selected arbitrarily. Phrases fall into natural categories like prefix, adjective, suffix, and many others. Even with the category properly established, the phrase choice is still context-dependent; hence the outcome may be risky. We therefore need to infer both the phrase location and the phrase itself–also by its type–at the same time, given enough contexts in the training or design phase.

It also seems obvious to us that creating rules manually–even for this apparently simple task–would either lead to oversimplification, or to a tedious work resulting in a solution hard to maintain. Moreover, we want the model to insert multiple phrases into the sequence in a flexible way. This motivated us towards developing the model we describe in the next section.

4. NEURAL MODEL DETAILS

We represent a neutral input sentence $\mathbf{u}$ as a sequence of tokens

$$\mathbf{u} = (w_1, ..., w_n, EOS), \quad w_i \in \mathcal{W},$$

where n is determined by the sentence length and may vary instance to instance. $\mathcal{W}$ denotes the vocabulary derived from a provided training data set, extended by the special token EOS used to mark the end of a sentence. We also add the UNK token to represent a word unknown to the model during training. It will be necessary in the inference phase discussed further.

We want our model to infer a family of probability distributions

$$P(X_i = p_j | \mathbf{u}, s, m; \theta), \qquad p_j \in \mathcal{P}$$

where X_i is a random variable for the phrase inserted at position $i \in \{0, ..., n\}$, over all phrases p_j observed in the training data set. The probabilities are conditioned over the utterance $\mathbf{u}$, sentiment s and its scalar-valued intensity (or *magnitude*) m. We define $\mathcal{P}$ as the set of all observed phrases, extended with a special symbol NIL meaning no phrase present at a given location. We index insertion points from zero to enable an insertion at the beginning of an utterance.

Our empirical distribution is determined by a training data set consisting of tuples

$$(\mathbf{u}, l, p, s, m),$$

where p is the (single) phrase inserted, $l \in \{0, ..., n\}$ its insertion point, s is categorical over predefined emotions, and, in our particular case, $m \in \{1, 2\}$ (corresponding to *moderate* and *intense* intensities).

From p and l we construct the ground truth matrix

$$\mathbf{P} \in \{0, 1\}^{\{0,...,n\} \times |\mathcal{P}|}$$

as follows: for position l

$$\mathbf{P}_{l,:} = \left[(\delta_j^k)_{j=1}^{|\mathcal{P}|} \right]^\top$$

where δ_j^k is the Kronecker delta (i.e. taking value 1 if $j = k$, and 0 otherwise), and k is the p phrase index in $\mathcal{P}$. So this is a one-hot distribution taking value 1 for the inserted phrase. For other locations, i.e. $i \neq l$, we define the corresponding slice $\mathbf{P}_{i,:}$ similarly, but taking the index of NIL as k. All probability mass is then concentrated on NIL, since there is no phrase in the particular location.

We use trainable word token embeddings and trainable phrase embeddings to reduce dimensionality and support generalization. We also represent the affective category as a one-hot vector with the intensity value. Hence the referred inputs become, respectively:

$$(\mathbf{w}_1, ..., \mathbf{w}_{n+1}); \quad \mathbf{w}_i = emb_w(w_i), \ \mathbf{w}_{n+1} = emb_w(EOS)$$
$$\mathbf{p} = emb_p(p),$$

and $\mathbf{s}$ being one-hot vector corresponding to emotion s, scaled further by m; where

$$emb_w : \mathcal{W} \to \mathbb{R}^{d_w}, \quad emb_p : \mathcal{P} \to \mathbb{R}^{d_p},$$

and $d_w, d_p \in \mathbb{N}$ are model hyperparameters.

Our model is a bidirectional Gated Recurrent Unit (GRU) network [15]. One GRU layer is fed $\mathbf{w}_{i+1} \circ \mathbf{s}$ each time step i (we index time steps from zero for further convenience, and we concatenate each token embedding with the sentiment vector being constant for each time step) until the end of the input sequence is reached. The layer generates intermediate hidden states $\mathbf{h}_i$. Another GRU layer is fed $\mathbf{w}_{n-i+1} \circ \mathbf{s}$ (i.e. going backwards from token $n+1$ down to 1), with generated hidden states denoted as $\mathbf{h}'_i$. Hidden states are further concatenated to create vectors $\mathbf{o}_i$.

$$\mathbf{o}_i = \mathbf{h}_i \circ \mathbf{h}'_{n-i}, \qquad i \in \{0, ..., n\}.$$

These vectors are then used to infer target phrase embeddings:

$$\widetilde{\mathbf{p}}_i = \mathbf{W}\,\mathbf{o}_i + \mathbf{b},$$

where $\mathbf{W}$ and $\mathbf{b}$ are trainable model parameters. Note that the dimensionality d_o of vectors $\mathbf{o}_i$ is fixed for all input data and for all i as it depends only on the model. We can then use the single matrix $\mathbf{W} \in \mathbb{R}^{d_p \times d_o}$ and the vector $\mathbf{b} \in \mathbb{R}^{d_p}$

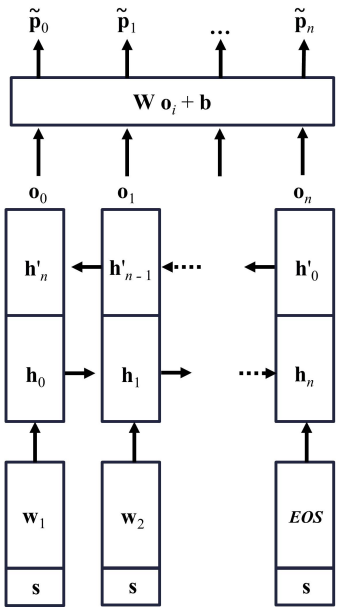

Figure 1. The model is fed with word tokens embeddings $\mathbf{w}_{i+1}$ ($\mathbf{w}_{n+1} = emb_w(EOS) =$ *EOS*) concatenated with the emotion vector **s**. Two GRU layers are used, the backward one is marked with apostrophes. Hidden states $\mathbf{h}_i, \mathbf{h}'_{n-i}$ are concatenated pairwise and undergo the affine transformation to match phrase embeddings space dimensionality. See Section 4 for detailed explanations.

for all affine transformations. Summarizing the above, we take hidden states pairwise from the two layers, concatenate them for every time step, and apply the affine transformation to match phrase embeddings' dimensionality. The model is presented in Figure 1.

During training we calculate scalar multiplications of each prediction $\widetilde{\mathbf{p}}_i$ by all phrases' embeddings $emb_p(p_j)$, apply the *softmax* function to translate the result to phrases' probabilities, and use the cross entropy H averaged over insertion points as the loss function L, all in the standard way [16].

$$\begin{aligned} \mathbf{logits}_i &= [emb_p(p_j)]_{j=1}^{|\mathcal{P}|} \bullet \widetilde{\mathbf{p}}_i, \\ \widetilde{\mathbf{P}}_{i,:} &= softmax(\mathbf{logits}_i)^\top, \\ L &= \frac{1}{n+1}\sum_{i=0}^{n} H(\mathbf{P}_{i,:}, \widetilde{\mathbf{P}}_{i,:}) + \alpha\|\mathbf{W}\|_F^2, \end{aligned}$$

the second loss term being the standard l_2 norm regularization weighted by the nonnegative hyperparameter α.

During inference we calculate $\widetilde{\mathbf{P}}$ from a given tuple $(\mathbf{u}, s, m)$. We then use a bunch of techniques to shape the final output. First, for every insertion point we take phrases with *softmax* probabilities exceeding some predefined threshold T and select a phrase randomly. This promotes output diversity. Second, if an inferred phrase is a neighbor of the *UNK* token, i.e. a word not observed during training, we can remove it to improve correctness. Last, we avoid generating phrases both at the beginning and at the end of a sentence. In such situation we choose one of them randomly to improve naturalness. All these steps are based on our empirical findings.

The model allows the situation where no phrase has been suggested for a given sentence, especially if we apply the limitations described above. In such case we can leave the phrase as-is, or e.g. retreat to a simple rule-based system. These options are discussed in detail in Section 6.

5. EXPERIMENTS

5.1. Training data

Our training data set was inspired by the Cornell Movie–Dialogs Corpus [17][1]. It contains utterances taken from movie scripts. We applied the DeepMoji classifier [18] to the data set and looked for pronounced affective categories to be used in our model. Using the classification we also selected a subset of utterances that we considered neutral. We then asked our linguist experts to add affective phrases to the selected sentences–at most one phrase for every sentence–covering all combinations of categories and intensities, trying to keep the original meaning. Emotions[2] we selected were *sadness*, *anger*, *doubt*, *happiness*, *affirmation* and *love*, with two intensities: *moderate* and *intense*. The resulting file contains 58,263 sentences with up to 8% variability between categories. This is due to the fact that we avoided adding phrases if the result might be considered unnatural. Here are some training set examples provided:

affirmation,1,Ah, coffee.< Okay.>
affirmation,2,<Absolutely! >She gets better.
anger,1,Make this <damn >Top Dollar smile.
doubt,1,All so sudden...< I don't get it.>
happiness,2,A call box?< How wonderful!>
love,1,I know<, sweetheart>.
sadness,2,A <helpless, >missing girl.

with numbers denoting emotion intensities (see the parameter m of the model in Section 4).

The resulting corpus contains 11,762 unique phrases, most of them assigned to *sadness* (2,256 phrases), and least to *affirmation* (1,816 phrases). Note that only some of them have been considered in the model training, as discussed in the next subsection. The minimum – median – maximum sequence length in the training set were 2, 7, 62, respectively.

5.2. Training and inference

We randomly split the resulting corpus to training (80%) and validation (20%) sets, and derived experimentally the following hyperparameters: $d_o = 256, d_w = 100, d_p = 20, \alpha = 0.01$, and batch size of 64. We used Adam optimizer [19] and dropout [20] with keep probability of 0.8. As a

1. The dataset is available at https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html (as of June 2018).

2. We refer to the affective categories as to "emotions" in the broader sense intended in [2], including "non-strictly rational aspects of the Hearer", like *doubt* or *affirmation*.

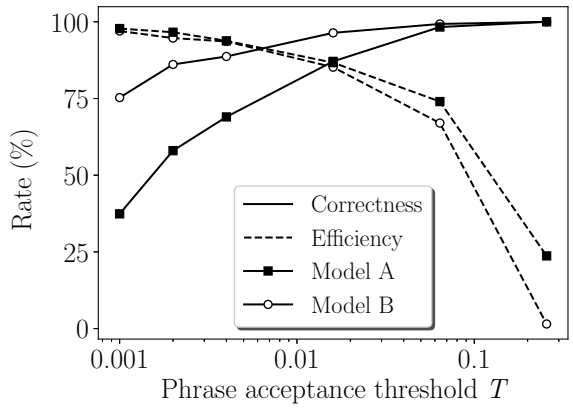

Figure 2. Tradeoff between model correctness and efficiency for various thresholds. For $T = 0.016$ the model B correctness was 96.4% with 85.2% efficiency. For $T = 0.064$ these became 99.3% and 74.0%, respectively. The horizontal scale on the plot is logarithmic. See subsection 5.2 for further details.

reference we trained a Seq2Seq model generating pairs of an insertion point and a phrase for an utterance, conditioned on emotion and intensity. Additionally we sampled the original train set to analyze inherent error rates.

To tune inference hyperparameters we manually created a new set of 84 neutral sentences distinct from the training data set. Then we inferred 1,008 affective utterances for each hyperparameters combination. We considered the model correctness defined using the minimum acceptance criteria for production use, that is, the grammatical validity and emotion matching; we disregarded the emotion intensity at this stage. The other measure of our interest was *efficiency*, defined as the rate of non-empty inferences.

We considered a grid of two hyperparameters: the softmax probability threshold T taking values 0.001, 0.002, 0.004, 0.016, 0.064, 0.256, and the training set truncation to 73 (model A) or 1,574 (model B) unique most popular phrases (corresponding to 35% and 80% of the training set, respectively, with 2,126 and 2,132 word vocabulary sizes). The selection of hyperparameter values was sparse due to the cost of the manual assessment by the linguistic panel. Results are presented in Figure 2. From them we can see that the model B performs better, most likely due to the bigger data set, and the tradeoff there between the model correctness and its efficiency. Certainly the numbers provide little guidance regarding the practical use because of the way the test set was created.

5.3. Test data

We prepared the final test set independently of the inference experiments mentioned above. To ensure objective representation we opted for a random selection of sentences from several corpora varying with domains and registers. We drew the samples in equal proportions from The EuroParl [21], British National Corpus [22], Manually Annotated Sub-Corpus of The Open American National Corpus [23], Santa Barbara Corpus of Spoken American English [24], and The Ubuntu Dialogue Corpus [25]. The set consists of sentences meeting the following general requirements:

1) Since the training set comprises first-person speaker utterances, 80% of the test set involves spoken language transcripts; only the Ubuntu Dialog Corpus is a written text.
2) Most of the test sentences are short, the median is 11 tokens. When longer than ten words, they were split into two if possible.
3) We tried to provide relatively neutral sentences devoid of content that might trigger emotional or judgmental associations.

We selected every tenth row satisfying the requirements. Due to formatting differences among corpora, a sentence originally split into several rows was treated as a whole. Incomplete sentences were omitted. The final set contains 100 utterances, resulting in 1,200 inferred sentences. A sample follows:

sadness, 1: This is an important matter. **I'm sorry!**
sadness, 2: **What a pity.** This is an important matter.
anger, 1: This is an important matter**, you fool**.
anger, 1: What is your **damn** question?
anger, 2: This is an important matter**, for fuck's sake**.
doubt, 1: This is an important matter. **I'm not sure about it.**
doubt, 2: This is an important matter **and I'm totally puzzled**.
happiness, 1: This is an important matter**, which is so nice**.
happiness, 2: This is an important matter **and I love it so much**.
affirmation, 1: **Okay.** This is an important matter.
affirmation, 1: **That's true.** It's yellow **and I think you're right**.
affirmation, 2: This is an important matter**, for sure**.
love, 1: This is an important matter**, my love**.
love, 2: This is an important matter**, my lovely sweetheart**.

The minimum – median – maximum sequence length in the test set were 2, 7, 62, respectively. Hence they differ substantially from the training set.

6. RESULTS

Only after selecting the final test set we ran the verification. To maximize the performance we trained the final model on 3,337 unique most common phrases, constituting 85% of the entire training set. Notably, the model was trained on a regular CPU (single quad-core Intel i7-6700 @3.40GHz was used) and converged to its locally optimal validation perplexity of 1.55 in just 30 minutes, producing 7.4 MB of resulting files. Therefore we speculate that the model should be robust to scaling to big data sets. Also, the training data set we worked on was limited in size, still giving us the very high correctness. We ran inference for the threshold $T = 0.02$ and $T = 0.06$, for all emotion–intensity combinations. We also used a binary parameter to include or exclude phrases neighboring tokens unknown during training. In case of an empty inference we had the option to keep the utterance unchanged, or use a simple rule-based model described further.

The model displayed the high correctness behavior from the very beginning, hence it was problematic to find a relevant automatic measure for it. Recent research [26] shows

TABLE 1. FINAL RESULTS FROM VARIOUS MODELS: BASELINE SEQ2SEQ, SIMPLE RULE-BASED, AND OURS: PURE DL TAKING ONLY NONEMPTY INFERENCES FROM THE DEEP LEARNING MODEL, AND HYBRID–WITH THE SIMPLE RULE-BASED SYSTEM APPLIED TO THE EMPTY ONES. WE ALSO SAMPLED THE ORIGINAL TRAINING SET TO ENABLE QUALITY COMPARISON. WE CONSIDER TWO PHRASE ACCEPTANCE THRESHOLD VALUES T. ALSO WE CONTROL IF WE KEEP (*on*) OR REMOVE (*off*) PHRASES NEIGHBORING TO WORDS UNKNOWN DURING TRAINING (OUT-OF-VOCABULARY, OOV) FROM THE RESULTING UTTERANCE. WE MANUALLY ASSESS GENERATION CORRECTNESS AND NATURALNESS; SEE SECTION 6 FOR METHODOLOGY DETAILS.

Model	T	OOV	Correctness (%)	Naturalness (%)	Efficiency (%)
`Training set`	-	-	95.6	99.2	-
`Seq2Seq`	-	-	65.8	56.8	-
`Rule-based`	-	-	81.5	58.3	-
`Hybrid`	0.02	*on*	79.4	72.1	91.9
		off	90.8	88.4	82.6
	0.06	*on*	83.5	71.7	79.3
		off	**94.9**	**88.4**	**100**
`Pure DL`	0.06	*off*	88.3	71.2	65.5

TABLE 2. ADVERSE EXAMPLES BY CATEGORY, TAKEN FROM THE MODEL THAT WE CONSIDER OPTIMAL; SEE SECTION 6 FOR DETAILS.

Invalid category (0.26%)
happiness, 1: The debate is closed**, honey**.
happiness, 1: They should not be about creating additional layers of bureaucracy and red tape**, sweetie**.
Invalid intensity (1.5%)
doubt, 1: It's yellow **and I'm totally puzzled**.
doubt, 1: Most of them. **I'm completely confused about the whole thing.**
affirmation, 1: Most of them. **No doubt about it.**
Invalid location (1.3%)
doubt, 1: **That's strange.** When it **probably** comes to safety my Group will always support any initiatives (...).
sadness, 1: **Sadly,** Or does she talk different because she's on the phone?
affirmation, 2: **Yes, of course!** You go **certainly** look, and every horse's hoof is shaped different.
Altered meaning (2.4%)
doubt, 2: Then we should follow the usual procedure, hearing one speaker in favor and one against. **It's impossible!**
doubt, 2: How's it gone today? **I have no idea!**
happiness, 2: Golly, I never really got into poetry. **It's so great!**
Output considered unnatural (11.6%)
love, 1: Madam President, on a point of order**, sweetie**.
sadness, 2: The debate is closed **and it hurts so much**.
love, 2: This latter point is of particular importance. **I simply adore it!**
Unintended irony (0.78%)
love, 1: **Sugar,** this latter point is of particular importance.
anger, 2: Well thank you very much and thank you for sharing your story with me **asshole**.
happiness, 2: I have installed qt3-designer now I don't know how to start it. **It's so great!**

that using automated measures like BLEU [27] or ROUGE [28] can be misleading for NLG performance evaluation. During training, validation was guided by the perplexity. However it turned out to be data set dependent: we reached 1.31 for the model A, and 1.50 for the model B. We then needed to turn to human evaluation.

We used the following methodology for the strict correctness assessment: each inferred set was evaluated independently by three linguists. We designated five obligatory binary criteria:

1) category: suitability for the specified affective category,
2) intensity: match for the intended level of intensity (*moderate* or *intense*),
3) location: syntactic correctness; the whole utterance was regarded as incorrect if at least one of the phrases disturbed the syntax,
4) meaning: absence of a significant semantic shift (e.g. *Coz it's all there, you know* vs. *Coz it's all there, you know**, you fool***), including unintended moves in deep-level structure (e.g. ***Sugar,*** *you don't need to count calories*),
5) naturalness: fluency and likelihood of occurrence in a casual conversation (e.g. number of phrases within a single utterance),

and two optional criteria:

1) other: less common phenomena worthy of note (e.g. repetitions or mismatched articles),
2) irony/sarcasm: emergence of ironic/sarcastic value, often being a decisive condition for acceptability.

We tried not to overspecify the criteria, because independent analysis reveals subtleties and uncertainties unsolvable without further assumptions. The judges were therefore asked to consult a list of the most common phrases categorized by emotion and intensity, if in doubt. For the final assessment we used a strict *correctness* demand as a logical conjunction of all 1–4 conditions, and considered *naturalness* as a separate dimension. The remaining criteria were considered

TABLE 3. HUMAN EVALUATION RESULTS PER CORPUS, FOR CORRECTNESS, NATURALNESS, AND UNINTENDED IRONY. THE CORPORA ARE THE EUROPARL [21], BRITISH NATIONAL CORPUS [22], MANUALLY ANNOTATED SUB-CORPUS OF THE OPEN AMERICAN NATIONAL CORPUS [23], SANTA BARBARA CORPUS OF SPOKEN AMERICAN ENGLISH [24], AND THE UBUNTU DIALOGUE CORPUS [25]

Corpus	% Correctness	% Naturalness	% Irony
[21]	93.4	83.8	2.6
[22]	**94.9**	**92.5**	**0.0**
[23]	92.2	92.5	0.7
[24]	**96.8**	**94.6**	**0.0**
[25]	95.9	78.6	0.6

informational. Since we needed binary partial values to assess the correctness, we used voting for each condition.

During our research we also introduced a simple rule-based system working by adding a random phrase in the prefix or suffix position. The phrase was selected from a small set of the most frequent phrases, up to 17 per a category/intensity combination. Initially it was only considered as a way to improve efficiency, but later on we discovered that the hybrid outperforms both the rule-based, and the pure DL model. This effect is somehow surprising and requires further research.

The results are presented in Table 1. For baseline comparison we took the abovementioned rule-based system, and a vanilla Seq2Seq model with attention. These models modify every utterance provided, so their efficiency is not reported. We also sampled the original training set and performed the assessment in order to measure the inherent error rate.

Our analysis shows that the hybrid model performs optimally with the acceptance threshold set to $T = 0.06$, and when we choose to avoid phrases located near words unknown during training. In this setup we reached 94.9% of the strict correctness and 88.4% naturalness with 100% efficiency, closely matching the correctness of the original training set (95.6%). Unintended irony or sarcasm was pointed out in six utterances (0.78% of all test cases). This adverse effect is typically caused by a clash of registers–e.g. combining the formal language with intense emotions–or from adding emotion to an utterance that is not neutral by its meaning. Because such situations cannot be completely avoided, we conclude that our model is matching human performance and is production ready. We provide detailed adverse examples per category in Table 2. This is visible in the dependency of the generation quality on the corpus selection, as demonstrated in Table 3.

7. CONCLUSION

We propose a deep learning model for the affective natural language generation task defined in the beginning, and we demonstrate its commercialization potential. In particular, the optimal model closely matches (by a margin of 0.7%) the human performance on the task, with applying the strict correctness criteria proposed in this paper. Additionally, the model was fast to train even on a general purpose CPU.

Admittedly, there are certain limitations resulting from the approach simplicity. Natural languages have got virtually unlimited ways of adding emotion to utterances, and many were out of our reach. Consider the following examples [18]:

(affirmation) This is **the shit**.
(anger) My flight is delayed.. **Amazing**.

This is–among other reasons–caused by the fact that we focus solely on tactical NLG choices, i.e. staying close to the surface realization, so we miss out the NLG strategy [29].

Another limitation we found is the tradeoff between the model correctness and its efficiency. We state that it is fair to allow the model to refrain from nonempty phrase generation, rather than produce a nonsense utterance. But we also observed that most–e.g. close to 90% in the model we consider optimal–of the empty inferences were related to existence of tokens unknown during training. This shows the potential to improve the efficiency by simply extending the training set, or by using pretrained embeddings. Interestingly, we also found that retreating to a simple rule-based system may actually improve the results.

Nonetheless we believe that the approach we propose is a promising solution to the DL precision issue. We hope that this publication will pave the way towards precise generative NLG methods. Our next plans are to investigate the model potential further: we would like to verify its applicability to other NLG tasks, like stylistic modifications with other expressive categories. Also we plan to experiment with other variations preserving the original sentence order, like augmenting the model with multipoint word replacements, or words and phrases deletion.

References

[1] Albert Gatt and Emiel Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.

[2] Fiorella De Rosis and Floriana Grasso, "Affective natural language generation," in *Affective interactions*, pp. 204–218. Springer, 2000.

[3] Ehud Reiter and Robert Dale, *Building natural language generation systems*, Cambridge university press, 2000.

[4] Michael Fleischman and Eduard Hovy, "Towards emotional variation in speech-based natural language processing," in *Proceedings of the International Natural Language Generation Conference*, 2002, pp. 57–64.

[5] Fazel Keshtkar and Diana Inkpen, "A pattern-based model for generating text to express emotion," in *Affective Computing and Intelligent Interaction*, pp. 11–21. Springer, 2011.

[6] Christina R Strong, Manish Mehta, Kinshuk Mishra, Alistair Jones, and Ashwin Ram, "Emotionally driven natural language generation for personality rich characters in interactive games.," in *AIIDE*, 2007, pp. 98–100.

[7] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[8] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou, "Affective neural response generation," in *European Conference on Information Retrieval*. Springer, 2018, pp. 154–166.

[9] Ondřej Dušek and Filip Jurčíček, "Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings," *CoRR*, vol. abs/1606.05491, 2016.

[10] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. Young, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," *CoRR*, vol. abs/1508.01745, 2015.

[11] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer, "Affect-lm: A neural language model for customizable affective text generation," *CoRR*, vol. abs/1704.06851, 2017.

[12] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba, "Sequence level training with recurrent neural networks," *CoRR*, vol. abs/1511.06732, 2015.

[13] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly, "Pointer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2692–2700.

[14] Jekaterina Novikova, Oliver Lemon, and Verena Rieser, "Crowd-sourcing nlg data: Pictures elicit better data," *arXiv preprint arXiv:1608.00339*, 2016.

[15] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. 2014, pp. 103–111, Association for Computational Linguistics.

[16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016.

[17] Cristian Danescu-Niculescu-Mizil and Lillian Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," in *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, 2011, pp. 76–87.

[18] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 1615–1625, Association for Computational Linguistics.

[19] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[20] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] Philipp Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, 2005, vol. 5, pp. 79–86.

[22] B Edition, BNC Baby, and BNC Sampler, "British national corpus," .

[23] Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau, "The manually annotated sub-corpus: A community resource for and by the people," in *Proceedings of the ACL 2010 conference short papers*. Association for Computational Linguistics, 2010, pp. 68–73.

[24] John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey, "Santa barbara corpus of spoken american english," *CD-ROM. Philadelphia: Linguistic Data Consortium*, 2000.

[25] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau, "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems," *arXiv preprint arXiv:1506.08909*, 2015.

[26] Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," *CoRR*, vol. abs/1603.08023, 2016.

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[28] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[29] Ielka van der Sluis and Chris Mellish, "Using tactical nlg to induce affective states: Empirical investigations," in *Proceedings of the fifth international natural language generation conference*. Association for Computational Linguistics, 2008, pp. 68–76.