Learning to Generate Examples for Semantic Processing Tasks

Danilo Croce
Dept. of Enterprise Engineering
University of Rome, Tor Vergata
Roma, Italy
croce@info.uniroma2.it

Simone Filice
Amazon
Tel Aviv, Israel
filicesf@amazon.com

Giuseppe Castellucci
Amazon
Seattle, USA
giusecas@amazon.com

Roberto Basili
Dept. of Enterprise Engineering
University of Rome, Tor Vergata
Roma, Italy
basili@info.uniroma2.it

Abstract

Even if recent Transformer-based architectures, such as BERT, achieved impressive results in semantic processing tasks, their fine-tuning stage still requires large scale training resources. Usually, Data Augmentation (DA) techniques can help to deal with low resource settings. In Text Classification tasks, the objective of DA is the generation of well-formed sentences that (*i*) represent the desired task category and (*ii*) are novel with respect to existing sentences. In this paper, we propose a neural approach to automatically learn to generate new examples using a pre-trained sequence-to-sequence model. We first learn a task-oriented similarity function that we use to pair similar examples. Then, we use these example pairs to train a model to generate examples. Experiments in low resource settings show that augmenting the training material with the proposed strategy systematically improves the results on text classification and natural language inference tasks by up to 10% accuracy, outperforming existing DA approaches.

1 Introduction

Deep Learning models achieve state-of-the-art results in many domains, including Computer Vision and Natural Language Processing (NLP). Training these large models typically requires many examples, whose collection and annotation can be costly and time-consuming. Data augmentation (DA) has proven an efficient way to acquire more training samples without incurring in the prohibitive annotation cost in a variety of fields, including computer vision (Perez and Wang, 2017) and speech recognition (Rebai et al., 2017). In NLP, some DA techniques have been proposed too, as surveyed in (Feng et al., 2021): common approaches create synthetic data by manipulating real examples, using Text Editing (Wei and Zou, 2019) or Back-Translation (Sennrich et al., 2015); the resulting examples are automatically labeled by inheriting the class of the original example they were generated from.

Unfortunately, when using recent pre-trained language models, such as BERT (Devlin et al., 2018) or RoBERTa (Liu et al., 2019), the effectiveness of DA methods is extremely limited, and sometimes they can even hurt the results (Longpre et al., 2020). A possible explanation for this inefficacy is that these DA techniques introduce lexical and structural variability that accurate language models directly induce during pre-training. The usefulness of synthetic examples is strictly related to their diversity from the original training data. At the same time, diverging too much from the initial data might increase the risk of introducing noisy annotations, i.e., synthetic data not reflecting the class of the original examples they were generated from.

To directly tackle the trade-off between diversity and label consistency, in this paper we propose DATS[1] (Data Augmentation based on Task-specific Similarity), a novel data-augmentation technique for text classification based on Natural Language Generation (NLG) models. Starting from a reduced set of annotated examples, we first learn a task-oriented similarity function that we use to automatically create pairs of similar examples. Then, we use these pairs to train a generative model to generate an example similar to the input one. Finally, we employ this model to generate new synthetic examples and augment the training data. We show that pairing examples with respect to their task-

[1] https://github.com/crux82/dats.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4587 - 4601
July 10-15, 2022 ©2022 Association for Computational Linguistics

oriented similarity is striking in order to allow the generative model to automatically understand the lexical and structural variations that can be applied to an instance without changing its label. Experimental results on four different text classification datasets demonstrate that DATS achieves better results than several existing DA solutions and that it systematically improves NLU models based on state-of-the-art pre-trained language models. In the remaining, Section 2 summarizes related works, Section 3 describes our method and Section 4 provides the experimental evaluation.

2 Related Work

Most of the existing approaches for DA perform some token-level manipulations on individual sentences (Kolomiyets et al., 2011; Kobayashi, 2018). For instance, Easy Data Augmentation (EDA) (Wei and Zou, 2019) applies simple operations including synonym replacement and random swap. These operations are also performed in (Ren et al., 2021) where the framework named Text AutoAugment (TAA) uses Bayesian Optimization algorithm to automatically search for the best manipulation policy. On the contrary, Wu et al. (2018) uses mask random tokens and use BERT to generate substituting words. Similarly, Kumar et al. (2020) use transformer-based models to apply token-level or span-level text content manipulation.

Other attempts operate at the entire sentence-level by paraphrasing the original text using back-translation (BT) (Edunov et al., 2018; Shleifer, 2019); however, BT tends to skew towards high-frequency words which not only causes redundancy but also leads to lexical shrinkage in the augmented data (Liu et al., 2020). More recent approaches for data augmentation use generative models to create more diverse synthetic data. Anaby-Tavor et al. (2019) fine-tuned a GPT-2 model to generate text given a target class, and use a text classifier to filter out those synthetic examples whose predicted class does not match the target class. In our work, we show that by conditioning a target model not only on the class labels, but also on representative examples, it is possible to better control the diversity-consistency trade-off.

3 Learning to generate examples

Recent advances in NLG (Vaswani et al., 2017) demonstrated that sequence-to-sequence (seq2seq) models can generate natural sounding and meaningful text given a prompt. As shown in (Keskar et al., 2019), the prompt can include style or content-related information that can help control the generation process.

In our work, we capitalize these techniques to augment a dataset $\mathcal{D}$ for the training of a text classifier $\mathcal{C}$. Specifically, our goal is to fine-tune a pre-trained seq2seq model $\mathcal{M}(*) = e$ that, given a prompt $*$, generates novel examples not in $\mathcal{D}$. The first application of this idea was investigated in Anaby-Tavor et al. (2019), where the authors trained a seq2seq model $\mathcal{M}(c) = s_o$ that generates examples s_o of a given class c. While this approach provides interesting results, it must be said that the variability of the generated examples can be quite low: by conditioning $\mathcal{M}$ only on c, it tends to produce the most frequent syntactic and lexical patterns associated with c, neglecting other modes this can exhibit. The key challenge here is to ensure diversity while preserving consistency, i.e., generating diverse examples that are valid representatives of the desired class.

We propose to train a seq2seq model $\mathcal{M}$ that is conditioned not only on the class c, but also on an example s_i of that class, i.e., such that $\mathcal{M}(c, s_i) = s_0$. The model is thus expected to synthesize a new example s_0, which is consistent with the input class c and is also "inspired" by an existing example s_i. The problem now is how to build the dataset to train $\mathcal{M}$. Pairing two random examples s_i and s_o belonging to the same class might be a viable solution. However, in classification tasks, examples are not necessarily similar, and coupling together radically different examples risks destabilizing the training of $\mathcal{M}$. Ideally, we would like to identify different modes in the same class and pair only examples belonging to the same mode. An alternative way to reach the same outcome is the adoption of a semantic similarity function that can be used to select linguistically related example pairs. Several unsupervised metrics exist (Croce et al., 2011; Cer et al., 2018; Poerner et al., 2020), however, they are inherently task-independent and not adequate in capturing task-specific similarities: two sentences such as "*The movie is good*" and "*The movie is not good*" (or even "*The movie is awful*") should not be paired together, when dealing with sentiment classification, while they are good candidate pairs in the training of a topic classifier.

To define a task-oriented similarity measure, we leverage the text classifier $\mathcal{C}$ we want to improve via data augmentation. Independently on the underlying neural architecture, a neural classifier has an encoder $\mathcal{E}$ that projects an example $s \in \mathcal{D}$ in a d-dimensional space, i.e., $\mathcal{E}(s) = \vec{s} \in \mathbb{R}^d$. In these spaces, simple linear classifiers (i.e., the output layers) identify the sub-spaces reflecting the target classes (Goodfellow et al., 2016; Goldberg, 2016). As a consequence, $\mathcal{E}$ is expected to project examples in sub-spaces representative for individual classes. Our task-aware similarity measure is the cosine similarity operating on these representations. For instance, after training a BERT model on the question classification task (Li and Roth, 2006), the encoding `[CLS]` of the entire question "*Who developed the vaccination against polio?*" allows retrieving other questions corresponding to the most similar `[CLS]` embeddings, such as "*Who invented the Moog Synthesizer?*": they are clearly not paraphrases, and their purely lexical similarity is quite low, but they show a similar pattern characterizing the question class `[HUM]` (i.e., human).

```
Input: A dataset $\mathcal{D}$, number of folds $f$, input pairs per example $n$ ,
       number of generated examples per input $l$ ,
Output: An augmented dataset $\mathcal{D}_{synt}$
$\mathcal{D}_{synt} = \emptyset$
$\mathcal{C} = $ TRAIN_CLASSIFIER$(\mathcal{D})$                    /* Train the classifier */
$\mathcal{E} = $ GET_ENCODER$(C)$                       /* Get the encoder used by $\mathcal{C}$ */
$\mathcal{D}_1, \ldots, \mathcal{D}_N = $ SPLIT_IN_FOLDS$(\mathcal{D}, f)$
for $i = 0$ to $f$ do
    $\mathcal{D}_{tr} = \bigcup_{\forall j \neq i} \mathcal{D}_j$           /* Split $\mathcal{D}$ in a $N$-cross fold schema */
    $\mathcal{D}_{te} = \mathcal{D}_i$
    $\mathcal{T} = \emptyset$       /* Initializing the training set $\mathcal{T}$ for the seq2seq model */
    foreach $e_i \in \mathcal{D}_{tr}$ do
        /* Focus only on examples having the same category of $e_i$          */
        $c_{e_i} = $ GET_CATEGORY$(e_i)$
        $\mathcal{D}^c_{tr} = $ SELECT_BY_CLASS$(\mathcal{D}_{tr}, c_{e_i})$
        /* Select the $n$ examples most-similar to $e_i$ in the embedding
           space generated by the encoder $\mathcal{E}$                       */
        $\mathcal{S} = $ TOP_SIMILAR$(e_i, \mathcal{D}^c_{tr}, \mathcal{E}, n)$
        /* Populate the training material for the seq2seq model $\mathcal{M}$     */
        foreach $e_o \in \mathcal{S}$ do
            $\mathcal{T} = \mathcal{T} \cup (c, e_i, e_o)$
        end
    end
    $\mathcal{M} = $ TRAIN_SEQ2SEQ$(\mathcal{T})$              /* Train the seq2seq model */
    foreach $e_s \in \mathcal{D}_{te}$ do
        $c = $ GET_CLASS$(e_s)$
        /* Generate $l$ examples from $e_s \in \mathcal{D}$ ignored in training $\mathcal{M}$ */
        $D_{synt} = D_{synt} \cup \mathcal{M}^l(c, e_s)$;
    end
end
return $D_{synt}$
```

Algorithm 1: Pseudo-code of DATS.

To generate an augmented dataset for a generic text classification task, we propose DATS (Data Augmentation based on Task-specific Similarity) that is described in Algorithm 1. First, we train a classifier $\mathcal{C}$ on $\mathcal{D}$. We use the resulting encoder $\mathcal{E}$ to project each training example e into the task-specific vector space, obtaining the embedding $\vec{e}$. Then, we split the training data $\mathcal{D}$ into $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$. Each example $e_i \in \mathcal{D}_{tr}$ is paired with the n examples $e_j \in \mathcal{D}_{tr}$ of the same class having the highest cosine similarity computed on their corresponding embeddings vectors $\mathcal{E}(e_i)$ and $\mathcal{E}(e_j)$. The resulting pairs are expected to lie in the same subspace and share some task-oriented linguistic relatedness. We use these example pairs to fine-tune a seq2seq model to solve the task $\mathcal{M}(c_{e_i}, e_i) = e_j$, where c_{e_i} is the category of e_i. Finally, examples in $\mathcal{D}_{te}$ are provided in input to $\mathcal{M}$ to generate the new synthetic dataset $\mathcal{D}_{synt}$. By applying multiple splits of $\mathcal{D}$ in a cross-fold scheme, we can use each training example to condition the model and generate new synthetic instances. As in almost all existing formulations, a generator can be used to generate a set of l variants for each input (c, e), for simplicity referred as $\mathcal{M}^l(c, e)$. In particular, techniques such as *nucleus-sampling* (Holtzman et al., 2020) enable the generation of large sets of variants, generally characterized by a good diversity.

It is worth noting that no assumption is applied when selecting $\mathcal{C}$ or $\mathcal{M}$. Even though in the experimental evaluation we consider only specific architectures (namely BERT and BART), this methodology can be applied to a wider plethora of models. Moreover, there is no restriction on the classification task type: the experimental section shows that DATS is applicable to classification tasks operating on both individual sentences and text pairs, e.g., in natural language inference (Bowman et al., 2015): given text pairs $s_{i,1}$ and $s_{i,2}$, it is sufficient to extend the above process defining $\mathcal{M}$ differently, i.e., $\mathcal{M}(c_{(s_{i,1}, s_{i,2})}, s_{i,1}, s_{i,2}) = (s_{j,1}, s_{j,2})$.

4 Experimental Evaluation

We test our approach on four tasks: 50-class Question Classification (QC) over the TREC dataset (Li and Roth, 2006); 5-class Sentiment Classification (SA) over the SST dataset (Socher et al., 2013); 7-class Intent Classification (IC) over the SNIPS dataset (Coucke et al., 2018); sentence-pair classification for Natural Language Inference (NLI) over the 3-category SNLI dataset (Bowman et al., 2015). Details about the datasets are in Appendix A.1.

Baselines. We compare our approach with multiple baselines and simplified versions of DATS for ablation study: (*i*) Easy Data Augmentation (EDA) (Wei and Zou, 2019); (*ii*) Back-Translation (BT) (Sennrich et al., 2015) in an English-German-English setting[2]; (*iii*) *Random Pairing* (RP) - DATS

[2] For EDA and BT we adopted the code in https://github.com/varunkumar-dev/

without task-oriented similarity functions where we created the training input-output pairs for $\mathcal{M}$ by selecting two random examples of the same class from the training set; (*iv*) Only Class (OC) - DATS where the prompt is only the category name (i.e., no representative example), similar to Anaby-Tavor et al. (2019).

Experimental Setting. For QC, SA and IC, we report the Accuracy when using q=10, q=50 or q=100 average examples per class[3]. For the NLI task, since it is more challenging, we report the Accuracy also for q=500 and q=1000 average examples per category. We also report the performance of each model when using the entire (F) training set[4]. We use the `bert-base-uncased` model from the Huggingface library (Wolf et al., 2019) as the classifier $\mathcal{C}$, and `bart-base` (Lewis et al., 2020) as the NLG model $\mathcal{M}$. Both are trained for 10 epochs with early stopping (patience=3) and learning rate $5e^{-5}$. We adopt nucleous sampling[5] (Holtzman et al., 2020) with p=0.90. We repeat the experiments 5 times with different seeds and we report the average classification accuracy.

Results. Table 1 shows the results of a BERT-based model without DA (NoDA), the DA baselines and our approach (DATS). We perform hyper-parameters tuning[6] on the development set of each task w.r.t. the number of similar examples n and the number of generated examples l. Task-agnostic DA approaches, like BT or EDA, seem slightly beneficial when using a transformer based classifier $\mathcal{C}$, as also reported in (Longpre et al., 2020). In some cases, they significantly hurt accuracy.

For instance, in QC or IC when q=10, EDA Accuracy is lower than NoDA by about 1 and 6 points, respectively. The same is for BT, where the drop is even higher, i.e., 8 and 18 points. Using the category information in the generation process, i.e., RP and OC, provides variable results, with minor improvements only in few specific settings (e.g., RP on SA). Instead, DATS improves accuracy for almost all q and tasks and such improvements are often statistically significant. When q=10, the improvement ranges between 1 and 10 points. In the few cases where a minor drop is observed, this is never statistically significant. By comparing DATS with RP it is clear that the task-specific similarity measure is striking to learn a good data generator[7]. DATS works well also on sentence-pair tasks, like NLI. However, given the complexity of this type of tasks, it requires more training examples (e.g., q>100) to show a positive impact. Furthermore, DATS outperforms other competitive methods [8]. With q=10 Kumar et al. (2020) report 67.30% accuracy in QC and 87.24% in IC, which is 1.74% and 4.5% lower than DATS. In the same setting, the gap with Wu et al. (2018) is even larger, i.e., 4.71% and 5.95%. Moreover, Ren et al. (2021) report 52.55% accuracy in SA using the entire training set, and DATS outperforms it by 1.71%.

q	**NoDA**	**EDA**	**BT**	**RP**	**OC**	**DATS**
Question Classification (QC)						
10	66.39	65.32	58.76*	66.72	62.64	**69.04**
50	90.64	90.48	89.84	90.28	90.48	**90.68**
100	91.60	92.04	91.48	91.08	91.44	**92.28**
F	91.60	91.32	91.48	90.72	91.36	**92.28**
Sentiment Analysis (SA)						
10	26.64	27.95	27.50	**29.82***	27.68	28.14
50	40.06	40.89	39.90	40.18	40.20	**43.48***
100	42.58	44.38	43.79	43.16	44.26	**46.30***
F	53.10	53.29	53.40	52.56	52.76	**54.26**
Intent Classification (IC)						
10	81.92	75.94*	63.71*	81.92	70.06*	**91.74***
50	95.18	95.14	94.69	95.42	95.28	**95.62**
100	95.38	95.54	95.57	95.54	**95.78**	95.62
F	97.32	97.80	97.42	97.96	98.00*	**98.12***
Natural Language Inference (NLI)						
10	**40.67**	39.42	36.07*	35.78*	36.30*	38.74
50	**49.32**	47.26	42.37*	43.58*	46.08*	48.40
100	57.78	**58.92**	56.23	50.64*	49.60*	58.36
500	71.48	72.23	73.23*	71.04	63.78*	**73.66***
1k	75.10	76.26	76.60*	75.86	75.48	**76.86***

Table 1: Accuracy of DA approaches: the best results are in bold while the results higher than NoDA are underlined. The $\star$ symbol indicates statistically significant differences (p<0.05) with respect to NoDA according to an unpaired T-test.

Qualitative Analysis. To better understand the advantages provided by DATS in the consistency-diversity trade-off, we report a qualitative evalua-

`TransformersDataAugmentation`

[3] To maintain the original class distribution we randomly sample $10c$, $50c$ or $100c$ from the original training set, with c being the number of classes. We ensure that at least one example per class is sampled.

[4] We omit results on the full dataset for NLI (made of more than $500,000$ examples) for which DA is not needed.

[5] In preliminary evaluations, we tested alternative decoding strategies, e.g., standard beam search and k-sampling. Overall, nucleous sampling was superior in terms of diversity and consistency, in line with the literature (Holtzman et al., 2020).

[6] An analysis of the hyper-parameters is in Appendix A.4

[7] We tried to substitute the task-oriented similarity with a cosine similarity on USE embeddings (Cer et al., 2018) - a general-purpose sentence encoder. Preliminary results on QC were in between DATS and RP, demonstrating that a task-oriented similarity is beneficial.

[8] Due to lack of budget, we did not run experiments on these models. The reported comparisons are based on the results reported by the authors.

cat.	Original Data	DATS	BT
Positive	*A colorful , vibrant introduction to a universal human impulse , lushly photographed and beautifully recorded .*	*This is a real film.*	*A colourful, lively introduction to a universal human impulse, lavishly photographed and beautifully captured.*
	A generous , inspiring film that unfolds with grace and humor and gradually becomes a testament to faith .	*Is in the midst of a monumental adventure, and I can't wait to see what the next one looks like.*	*A generous, inspiring film that unfolds with grace and humor and gradually becomes a witness of faith.*
	A solid and refined piece of moviemaking imbued with passion and attitude .	*It 's a film that develops with conviction and a deep, passionate heart.*	*A solid and refined piece of film steeped in passion and attitude.*
	Enormously enjoyable , high-adrenaline documentary .	*It is a remarkable film...*	*Enormously entertaining, adrenaline-pumping documentary.*
Negative	*- spy action flick with Antonio Banderas and Lucy Liu never comes together .*	*" It just felt... too heavy... for me to have it in my trunk."*	*- Espionage action with Antonio Banderas and Lucy Liu never comes together.*
	If you value your time and money , find an escape clause and avoid seeing this trite , predictable rehash .	*Isn't the best spy film ever made.*	*If you value your time and money, find an exit clause and avoid this mundane, predictable repetition.*
	No amount of burning , blasting , stabbing , and shooting can hide a weak script .	*It is a masterpiece, brilliantly directed, and incredibly well done.*	*No amount of firing, blasting, stabbing and shooting can hide a weak script.*
	This Tuxedo ... should have been sent back to the tailor for some major alterations .	*This is a film that I was unable to stomach... and very much enjoyed... until the last 30 minutes... when I finally got close to the film... and it turned into a garbage dump.*	*This tuxedo... should have been sent back to the tailor for some major changes.*

Table 2: A qualitative analysis of the diversity in the Sentiment Analysis task over SST-5.

tion of the generated sentences for the SA task.

In particular, Table 2 reports some of the SA examples generated by different models on classes 5 and 1 when using $q = 10$ average training examples per class. Exception for a single case "(*It is a masterpiece, brilliantly directed, and incredibly well done.*)" that is not a negative sentiment example as it was supposed to be), the examples generated by DATS (with $n = 1$ and $l = 1$) are all label-consistent. However, the diversity introduced by DATS is impressive: all the examples are very different from the input example used to condition the NLG model while generating them, and therefore can be very useful when augmenting the training data of NLU models. On the opposite, BT introduces very minor modifications to the input text, resulting in a significant lower diversity A larger and systematic qualitative analysis on the quality and diversity of the generated material is in Appendix A.2 and A.3.

5 Conclusions

This paper proposes DATS, a data augmentation method based on Natural Language Generation (NLG) models. A generative model is fine-tuned to produce examples similar to the input ones. The training input-output pairs are selected according to a task-oriented similarity function. This pairing allows the NLG model to learn the lexical and structural variations that can be applied to an instance without changing its label. The experimental results suggest that the generated sentences are diverse and label consistent, and can improve state-of-the-art text classifiers, outperforming existing DA methods. In the future, we plan to apply DATS to further tasks (e.g., Question Answering) and neural architectures.

Acknowledgments

We would like to thank the "Istituto di Analisi dei Sistemi ed Informatica - Antonio Ruberti" (IASI) for supporting the experimentations through access to dedicated computing resources.

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2019. Not enough data? deep learning to the rescue! *CoRR*, abs/1911.03118.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces.

Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2011. Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1034–1046, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for NLP. *CoRR*, abs/2105.03075.

Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *J. Artif. Intell. Res.*, 57:345–420.

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge, MA, USA. `http://www.deeplearningbook.org`.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *CoRR*, abs/1805.06201.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 271–276, Portland, Oregon, USA. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xin Li and Dan Roth. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3):229–249.

Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9031–9041, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shayne Longpre, Yu Wang, and Christopher DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? *CoRR*, abs/2010.01764.

Luis Perez and Jason Wang. 2017. The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. Sentence meta-embeddings for unsupervised semantic textual similarity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7027–7034, Online. Association for Computational Linguistics.

Ilyes Rebai, Yessine BenAyed, Walid Mahdi, and Jean-Pierre Lorré. 2017. Improving speech recognition using data augmentation and acoustic model fusion. *Procedia Computer Science*, 112:316–322. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.

Shuhuai Ren, Jinchao Zhang, Lei Li, Xu Sun, and Jie Zhou. 2021. Text AutoAugment: Learning compositional augmentation policy for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages

9029–9043, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *CoRR*, abs/1511.06709.

Sam Shleifer. 2019. Low resource text classification with ulmfit and backtranslation. *CoRR*, abs/1903.09244.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR*, abs/1901.11196.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2018. Conditional BERT contextual augmentation. *CoRR*, abs/1812.06705.

A Appendix

A.1 Task Description

In this section we report details and statistics of the dataset we adopted in the experimental section.

TREC. The TREC (Li and Roth, 2006) dataset contains 4,907, 545 and 500 examples for training, development and test, respectively. We adopted the fine-grained version of the dataset that contains 50 categories.

SST5. The Sentiment Analysis Treebank dataset (Socher et al., 2013) consists of 8,544, 1,101 and 2,210 examples for training, development and test, respectively. The dataset is characterized by 5 categories for sentiment, i.e., 1 (Very Negative) 2 (Negative), 3 (Neutral), 4 (Positive) and 5 (Very Positive) and it contains movie review sentences.

SNIPS. The SNIPS dataset (Coucke et al., 2018) consists of natural language commands for a voice assistant. The commands are classified into 7 categories, i.e., RateBook, BookRestaurant, AddToPlaylist, PlayMusic, GetWeather, SearchScreeningEvent, SearchCreativeWork. The dataset consists of 13,084, 700, 700 examples for training, development and test, respectively.

SNLI. The SNLI dataset (Bowman et al., 2015) consists of examples of pairs for the Natural Language Inference task. There are 3 categories: *entailment*, *neutral* and *contradiction*. The dataset consists of 550,152, 10,000 and 10,000 for training, development and test, respectively.

A.2 Qualitative Analysis

This section reports additional examples generated by DATS. In particular, we report the examples generated with $n = 2$ and $l = 10$. Notice that, the generated examples are not in the original datasets. It is worth noting that some examples may result odd, such as a question like "*Name the American president born in 2005?*" but we only expected that they are linguistically sound and consistent with the corresponding class. In general, the number n strongly affects the example novelty: high values of n make the generator observe multiple times the same text as the target sentence, making the generation process more conservative. As a result, when using $n > 20$ the generator tends to produce the same examples from the original dataset.

Question classification. Here is a list of 20 questions generated by DATS for different classes. Please refer to the original dataset description for the details about the specific classes.

- `ABBR:exp` *What does D-DAY stand for?*
- `DESC:def` *What does the term "Italian Renaissance" mean?*
- `DESC:desc` *What process takes place after a hydrogen release?*
- `DESC:reason` *Why do people go to the bathroom at night?*
- `ENTY:animal` *What animal was the first domesticated creature in the world?*
- `ENTY:color` *What color is the cross on the French flag?*
- `ENTY:event` *What happened on February 27, 1991?*
- `ENTY:food` *What cereal is "sweet, soft and orange"?*
- `ENTY:plant` *What tree has the longest trunk?*
- `ENTY:subst.` *What are diamond rings made of?*
- `ENTY:veh` *What was the name of the U.S. Navy gunboat used by Dwight D. Eisenhower?*
- `HUM:gr` *Name the two major companies in the energy industry.*
- `HUM:ind` *What's the name of the author of "Harry Potter"?*
- `LOC:city` *What's the name of the largest city in Germany?*
- `LOC:country` *What country in 1991 recorded the largest number of cocaine seizures?*
- `NUM:count` *How many hundred ships sank in the Norwegian Sea in 1923?*
- `NUM:date` *What is the date of the first inauguration of President Nixon?*

- NUM:money *What is the cost of university admission to Stanford?*
- NUM:period *How long does it take to clean up a cache?*
- NUM:weight *What is the maximum weight for a healthy adult?*

Although the above examples seem coherent, we can argue if they can be really new, i.e., useful for the task according to a Data Augmentation process. We thus focused on the specific class ENTY:lang that is underrepresented in the original dataset, with only the six original examples reported below:

- *What is the name of a language spoken by the Sioux?*
- *What is the only modern language that capitalizes its singular first-person pronoun?*
- *What is one of the languages spoken by the Sioux called?*
- *What's the official language of Algeria?*
- *What are the two languages of Malta?*
- *What is the main language of Sao Paulo , Brazil?*

Below some of the synthetic examples generated by DATS with q=10. It is worth noting that language model introduces some expressions like *sub-dialect* that are novel with respect to the original training material.

- *What is the language of Switzerland?*
- *What is the oldest language in the Americas?*
- *What's the language spoken by the Kootenai people?*
- *What is the sub-dialect of English?*

Conversely, if we only use the category name as a prompt (i.e., the OC model), the generated questions will have a lower quality. For instance, with $q = 10$ and $l = 10$, the OC model produces questions like:

1. *What is the name of the island of Cote D 'Azubis?*
2. *What English language does the language speak?*
3. *What languages of English and French are spoken as well as Arabic?*
4. *What what languages would French be spoken?*

They have several issues, including label-inconsistency (question 1), malformed syntax (question 4) and unclear semantics (question 2-3).

Similarly, the questions generated by EDA seems of lower quality. For example, with $q = 10$, the EDA model produces the questions:

1. *What is the name of a language spoken by?*
2. *What is the name of a language spoken by the Sioux?*
3. *What is language mostly spoken in Brazil?*
4. *What language is mostly spoken in Brazil?*
5. *Name a gaelic language?*
6. *What is one of the spoken communication spoken by the Sioux called?*
7. *What is one of the languages spoken by the Sioux called?*
8. *What's the most commonly verbalise language in Belgium?*
9. *What's the most commonly spoken language in Belgium?*

Similarly, to the OC case, these questions have several problems, including malformed syntax (question 1, 3, 8), incompleteness (question 1).

Sentiment Analysis. Below is a list of statements generated by DATS in the movie review domain. To simplify the qualitative analysis, we report only Negative (class 1 in the dataset), Neutral (class 3) and Positive (class 5) examples. In general, the method generates sound judgments, with a high syntactic variability even though a quite limited lexical variability, i.e., a manual inspection suggested that most of the adjectives used to express judgments are reused in the generated examples. Sentences in general are less complex than the original ones. This is reflected by generally shorter generated sentences, i.e., with an average length of 16 words per review against the original 20 words per review.

`Negative` reviews:

- *It's almost too derivative to stand on its own as a stark portrait of desperation and violence.*
- *A mediocre work of storytelling, lacking the slightest bit of wit or charm.*
- *A self-conscious, incoherent, self-interested fable.*

`Neutral` reviews:

- *It could be a lot worse if it were, well, more adventurous.*
- *If Tuck had ever made a movie about a vampire, it probably would look a lot like this one.*
- *It's very much like George Romero's final work, where he had a hand in making huge cuts to his movie – only it takes a whole lot more to feel good about.*

`Positive` reviews:

- *A gripping and wildly unpredictable comedy.*
- *It's lovely, funny, different, odd.*
- *A compelling example of why animation is a part of the human spirit.*

The sentences generated using only the category as a prompt (i.e., OC model) have a lower quality. For instance, with the same q=10 the OC model sentences look like:

`Negative` reviews:

- *Why, a good movie, an interesting and absorbing film*
- *The, says a filmmaker and the film 's creator.*
- *The its full potential.*

`Neutral` reviews:

- *No at its most extreme.*
- *This and an epic comedy of a documentary.*
- *So a little razzle dazzle, but not at all riveting,.*

`Positive` reviews:

- *No the greatest of the time*
- *This the first movie to touch a dark heart*
- *No in the form of a poem and a movie itself*

Similarly, the sentences generated by EDA contain different issues, which can be explained by the substitution and word swapping strategies adopted by this method. For example,

- *This tuxedo should have been sent back to the tailor for john major some major alterations*
- *If you value your time and money find an escape clause rehash avoid seeing this trite predictable and*
- *Spy litigate flick with Antonio Banderas and Lucy Liu never comes together*

`Neutral` reviews:

- *It would work a better as much one hour tv documentary*
- *The fast runner transports the viewer into an unusual space*
- *Often overwrought and at times positively irritating the film turns into an engrossing thriller almost in spite of*

Positive reviews:

- *Documentary enjoyable high adrenaline enormously*
- *Quite simply a joy to watch follow and especially to listen to*
- *A solid and refined piece of moviemaking imbued with passion and attitude*

Intent Classification. Below is a list of examples generated by DATS of each intent in the SNIPS dataset. Please refer to the original dataset description for the details about the specific classes. In general, the examples are different from the original ones introducing variability mostly on the involved named entities, e.g., proper nouns of the music authors or places. In general, the syntactic complexity is the same as the original material. Sometimes, odd dates are introduced (here 2037 as a date for a reservation).

- AddToPlaylist *add this track to my modern psychedelia playlist*
- BookRestaurant *book a diner for 1 on feb 28th 2037*
- GetWeather *will it be warm in michigantown at 07:00:00 am*
- PlayMusic *i want to listen to the last track by michael hayvoronsky*
- RateBook *i give the current essay a four*
- SearchCreativeWork *find a painting called the night owl*
- SearchScreeningEvent *show me the movie schedule for national tv*

Conversely, using only the category as input to the generation model, i.e., the OC baseline, is not able to produce high quality examples. For instance, with q=10, the OC model generates sentences like:

- AddToPlaylist *Add please rewind now my playlist*
- BookRestaurant *Book the menu for the night in the hotel room*
- GetWeather *Add a weather forecast from my backyard*
- PlayMusic *Play tunes and tracks to the chino sound bar*
- RateBook *Add some jazz music*
- SearchCreativeWork *Add my novel*
- SearchScreeningEvent *Add movies this week*

Again, EDA is only able to generate some minor variations of the training examples, and sometimes the swapping/substitution strategies are introducing issues. For example:

- AddToPlaylist *I want this record album on my indie alternative playlist*
- BookRestaurant *I need a table for during midday in Montana*
- GetWeather *Is it going tea be freezing at to time in Michigantown KS*
- PlayMusic *Play any song from rebekah hewitt*
- RateBook *Rate this series one stars*
- SearchCreativeWork *The me show song spiderman of the rings*
- SearchScreeningEvent *I want the neediness movie schedule for animated movies in the area*

Natural Language Inference. This task is the most challenging as the generated text is not only expected to be internally consistent, but also pair-wise consistent. In general, the syntactic complexity of the sentences is preserved, with an average length of 10 words per sentence. Similarly to the original training set, in many entailment examples generated by DATS the premise and the hypothesis share the same verb (e.g., *watching* in the first example). On the contrary, the subject of the action is often changed during the contradictions (e.g., *boy* vs *girl*) in the first contradictory pair. Below we report some example pairs:

Text Pairs in Entailment

- "*Two men watching a sports event in the background.*" entails "*Two men are watching a game.*"
- "*A young woman, wearing sunglasses, is raising her hand.*" entails "*A woman is raising an arm*"
- "*A boy dressed to play soccer is playing" soccer.* entails "*The boy is playing outside.*"

Text Pairs in `Contradiction`

- "*Young Indian boy playing cricket and soccer.* contradicts "*Young Indian girl playing cricket.*"
- "*A boy in a yellow shirt, orange and white, playing guitar.*" contradicts "*A boy is sitting quietly.*"
- "*A man is attempting a jump on his skateboard.*" contradicts "*A man rides his bike over an obstacle course.*"

Instead, the examples generated when only prompting the model with the category name are not as good. For example, the OC model with *q*=10 produces the following examples:

Text Pairs in `Entailment`

- "*A man being photographed in a store in Seattle*" entails "*A man is sitting in a car.*"
- "*A woman leaves a beer in the air.*" entails "*A woman is leaving a beer.*"
- "*A man walks along a street in St. Pete and walks past a city bus terminal.* entails "*Police are standing near the ground holding a man*"

Text Pairs in `Contradiction`

- "*A white van is driving by and a white man is sitting next to it* contradicts "*A man in a blue polo shirt is sitting on the back of a white car.*"
- "*from the middle of a street at a supermarket street, two girls wearing black and a yellow shirt.*" contradicts "*A woman wearing a pink shirt looks down at a pole, waving a dollar bill.*"
- "*s a little boy playing a ball.*" contradicts "*A dog is jumping out of a field.*"

Notice that the OC model is, in general, not able to produce label consistent pairs. For example, the first and third pairs reported above for the entailment category are not correct examples of this class.

In the following, we report also some examples generated with EDA. Again, EDA strategies introduces some issues in the sentences. For example,

Text Pairs in `Entailment`

- "*An old man with gray hair wearing a scarf and black jacket*" entails "*An old woman is bundled up*"
- "*Two ladies are maam laughing on the street*" entails "*There are women*"
- "*Lady into the sky on city* entails "*Lady looks into the sky*"

Text Pairs in `Contradiction`

- "*Young man doing a trick jump on a skateboard*" "*A man is writing record*"
- "*A by holds hands while walking couple buildings*" contradicts "*A couple dance in the street*"
- "*A a man in a black shirt and shorts sitting at blond table wine with a glass of eating*" contradicts "*a woman is drinking a beer*"

Besides introducing grammatical errors, EDA is also introducing label consistency errors (for example, the first entailment pair).

A.3 Are generated examples really different from the original ones?

To better understand the advantages provided by DATS in the consistency-diversity trade-off, we here report an exhaustive qualitative evaluation.

Table 3 reports all the SA examples generated by different models on classes 5 and 1 when using $q = 10$ average training examples per class. Exception for a single case "(*It is a masterpiece, brilliantly directed, and incredibly well done.*)" that is not a negative sentiment example as it was supposed to be), the examples generated by DATS (with $n = 1$ and $l = 1$) are all label-consistent. Conversely, many examples generated with the OC baseline have a wrong sentiment. On the other side, the diversity introduced by DATS is impressive: all the examples are very different from the input example used to condition the NLG model while generating them. At the same time, they are also very novel with respect to the full training set (used to train the NLG model) and therefore can be very useful when augmenting the training data of NLU models. On the opposite, BT and EDA introduce very minor modifications to the input text, resulting in a significant lower diversity.

Table 4 provides examples of generated questions for the QC task. This task is particularly different from SA as it involves 50 categories and the size of the training dataset causes many of them to be underrepresented. We generated a set of questions with DATS ($q = 100$, $n = 1$ and $l = 1$) by selecting one instance for each category as the input to the model. Each row of the table thus reports the class and example used in input to the generator and the last column reports the generated example.

As it can be noticed, DATS generates examples different from the input ones: only 4 out of 50 (8 %) examples are exactly copied and 3 out of 50 (6%) are different with respect to only one word. Notice that the QC task is particularly sensitive to the syntactic structure of the sentence and this is generally preserved in the generation process. DATS very rarely generates a paraphrase of the original sentence. The target of the generated question almost always changes (e.g., when using ”*Who is Nicolo Paganini?*“ as input, DATS generates ”*Who is Michael Jackson?*“). Finally, notice that in all cases DATS is able to generate a question which is coherent with the target category of the input question.

category	**Input**	**DATS**
`ABBR:abb`	*What does the number ” 5 ” stand for on FUBU clothing?*	*What does the letter C stand for on a Trivial Pursuit board?*
`ABBR:exp`	*What is HTML?*	*What is video?*
`DESC:def`	*What are manifest and latent function theories?*	*What are endometriosis theories ?*
`DESC:desc`	*What ’s the origin of the word ’ news ’?*	*What is the origin of the word ’community ’?*
`DESC:manner`	*How do you build a solar car for a high school experiment?*	*How do you make a computer chip ?*
`DESC:reason`	*What causes someone to be ticklish?*	*What causes a person to have dry hands?*
`ENTY:animal`	*What animal migrates the farthest?*	*What animal scares people most?*
`ENTY:body`	*What are the names of the different toes?*	*What is the name of the sea where the sharks swim?*
`ENTY:color`	*What color is Mr. Spock ’s blood ?*	*What color are the veins in the neck of the elephant?*
`ENTY:cremat`	*Name the only extant trilogy of classical Greek plays .*	*What is the title of the new play, The Muppets, by Tom Wolfe?*
`ENTY:curren.`	*What type of currency is used in China?*	*What currency is used in Dominica?*
`ENTY:dismed`	*What therapy attempts to elicit the ” primal scream ”?*	*What pharmacotherapy methods work with patients with fibromyalgia?*
`ENTY:event`	*What famed gangland slaughter cut into the membership of Bugs Moran ’s outfit?*	*What was the outcome of the Main Conference?*
`ENTY:food`	*Madonna hocks which soft drink?*	*Which soft drink did Madonna advertise for?*
`ENTY:instru`	*Ray Charles is best known for playing what instrument?*	*Ray Charles is known for playing what instrument?*
`ENTY:lang`	*Name a Gaelic language .*	*What is the Gaelic language?*
`ENTY:letter`	*What ’s the second-most-used vowel in English?*	*What ’s the longest letter in the alphabet of a C major?*
`ENTY:other`	*What impenetrable system of French fortifications was built along the German frontier before World War II ?*	*What limestone wall supports the Great Wall of China?*
`ENTY:plant`	*What are two plants that clothes are made from?*	*What plants are found in the forest?*
`ENTY:product`	*What is the most advanced handheld calculator in the world?*	*What ’s the most efficient way to use a computer monitor?*
`ENTY:relig.`	*What Caribbean cult did Marcus Garvey originate?*	*What religion claims to control all nations?*
`ENTY:sport`	*The second most popular sport worldwide is what?*	*What is the most popular sport in the world?*
`ENTY:subst.`	*What are the ingredients of Coca-Cola?*	*What are the secretions of the Earth ’s crust?*
`ENTY:symbol`	*What is the zodiac sign for August 14?*	~~*What is the zodiac sign for August 14?*~~
`ENTY:tech.`	*What is the best way to travel in Japan?*	*What are the best ways to get a text message when you have a web address?*
`ENTY:termeq`	*How do you say I love you in Spanish?*	*How do you say I love you in French?*
`ENTY:veh`	*Which type of submarine was bought recently by South Korea?*	*What type of sub was the Mercury Project Explorer?*
`ENTY:word`	*What are the first six words of Dickens ’s A Tale of Two Cities?*	*What are the first words of Dickens ’s The Scarlet Letter?*
`HUM:desc`	*Who is Nicolo Paganini?*	*Who is Michael Jackson?*
`HUM:gr`	*What stereo manufacturer is ” Slightly ahead of its time ”?*	*What company produces’Prussia ’ ?*
`HUM:ind`	*Who was the first black woman to star in the Folies Bergeres?*	*Who was the second woman to star in the movie ” Logan ’s Littlest Baby ”?*
`HUM:title`	*What is her profession?*	*What is her occupation?*
`LOC:city`	*What is the largest city in the world?*	~~*What is the largest city in the world?*~~

LOC:country	*What two European countries are joined by the Gran San Bernardo Pass?*	*What two European countries share the Khyber Pass?*
LOC:mount	*Where do people mountain climb in Nepal?*	*Where do most people live?*
LOC:other	*What attracts tourists to Reims ?*	~~*What attracts tourists to Reims ?*~~
LOC:state	*What states do not have state income tax?*	*What state does not have a lottery?*
NUM:code	*What is Columbia Tristar 's phone number?*	*What is Judy Garland 's phone number?*
NUM:count	*How many corners does a spritsail have?*	*How many sides does a lobster have?*
NUM:date	*When did the Berlin Wall go up?*	~~*When did the Berlin Wall go up?*~~
NUM:dist	*How wide is the Atlantic Ocean?*	*How wide is the Peloponnesian Sea?*
NUM:money	*What is average salary of restaurant manager in United States?*	*What is the average salary of a military major?*
NUM:ord	*What chapter of Gone with the Wind has Rhett Butler leaving Scarlett O 'Hara?*	*What chapter of Gone with the Wind ends with the line : " To your parents or not to your parents. "?*
NUM:other	*What is the latitude and longitude of El Paso , Texas?*	*What is the population of San Diego, California?*
NUM:perc	*Of children between the ages of two and eleven , what percentage watch " The Simpsons "?*	*What percentage of Americans watch MTV?*
NUM:period	*What is the time it takes a typist to type a screenplay that is 100 pages long?*	*What is the time it takes the typical person to go to the bathroom?*
NUM:speed	*What is the speed of the Mississippi River?*	*What is the speed of a Corvette ?*
NUM:temp	*How hot should the oven be when baking Peachy Oat Muffins?*	*How hot should a chef cook dinner?*
NUM:volsize	*How big is our galaxy in diameter?*	*How big is the Moon?*
NUM:weight	*What is the weight of air?*	*What is the weight of a teaspoon of matter?*

Table 4: Examples of DATS outputs generated sentence on the QC task for each category.

A.4 Effects of Hyper-parameters

This section reports an analysis of the role of DATS hyper-parameters. In particular, we show the results by varying (i) the number of most similar elements n used to generate the pairs for training $\mathcal{M}$ and (ii) the number of elements l generated with nucleous sampling. Specifically, we tried $n \in [1, 2, 5]$ and $l \in [1, 2, 3, 5, 7, 10]$. In figure 1 we report the difference in Accuracy between our approach and the NoDA baseline for each specific configuration. Each figure refers to a specific q (i.e., the average number of examples per class) value, and reports the average delta accuracy computed on the QC, SA and IC tasks.

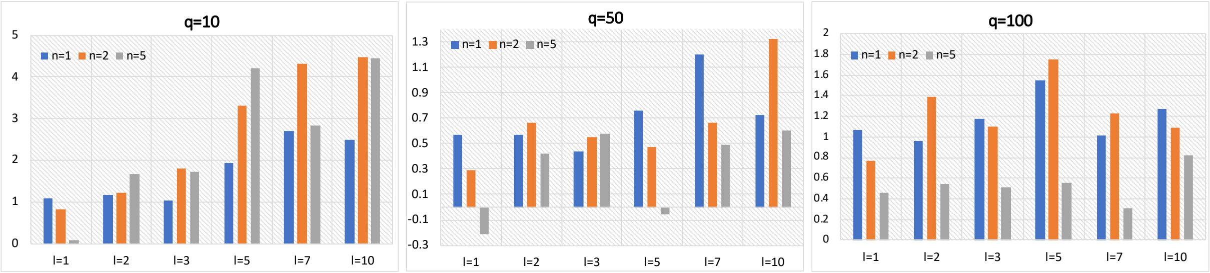

Figure 1: Average Accuracy difference between DATS and the NoDA baseline w.r.t. different dataset sizes. We vary the number n of pairs created with the learned similarity function and the number l of generated examples with generation model $\mathcal{M}$. The average is computed w.r.t. the QC, SA and IC tasks.

In general, when using small datasets it is beneficial to use higher n values, i.e., to create more pairs for training the NLG model. In fact, when q=10 we can observe that the best improvement is obtained with n=2 or n=5.

Regarding the number l of examples generated through $\mathcal{M}$, we can observe that generating a higher number of examples seems to be beneficial in almost every case, especially when dealing with a small dataset. For example, when q=10 or q=50, our DA approach provides the best performance with $l = 7$ or l=10. More generally, we can observe a positive trend in generating at least l=5 examples with the generation model.

cat.	Original Data	DATS	BT	EDA	OC
Positive	*A colorful , vibrant introduction to a universal human impulse , lushly photographed and beautifully recorded .*	*This is a real film.*	*A colourful, lively introduction to a universal human impulse, lavishly photographed and beautifully captured.*	*a colorful vibrant introduction to a universal human impulse man lushly photographed and beautifully recorded*	*No what a story is being.*
	A generous , inspiring film that unfolds with grace and humor and gradually becomes a testament to faith .	*Is in the midst of a monumental adventure, and I can't wait to see what the next one looks like.*	*A generous, inspiring film that unfolds with grace and humor and gradually becomes a witness of faith.*	*a generous inspiring film that unfolds with grace and humor and becomes a testament to faith*	*The 's most engaging and visually arresting show in its time of adventure and mystery.*
	A solid and refined piece of moviemaking imbued with passion and attitude .	*It 's a film that develops with conviction and a deep, passionate heart.*	*A solid and refined piece of film steeped in passion and attitude.*	*a solid and refined piece of moviemaking diffuse with passion and attitude*	*If is the best film yet written.*
	Enormously enjoyable , high-adrenaline documentary .	*It is a remarkable film...*	*Enormously entertaining, adrenaline-pumping documentary.*	*documentary enjoyable high adrenaline enormously*	*The film is a bit of a film of a sort.*
	Lan Yu is a genuine love story , full of traditional layers of awakening and ripening and separation and recovery .	*The film 's tone, texture, and storytelling are beautifully articulated.*	*Lan Yu is a true love story full of traditional layers of awakening and maturing and separation and recreation.*	*lan yu is a genuine love story full of traditional layers of awakening separation and recovery*	*As is a good story.*
	Quite simply , a joy to watch and – especially – to listen to .	*This is a human impulse, a sublime gift.*	*It's just a joy to watch and, above all, to listen.*	*quite simply a joy to watch follow and especially to listen to*	*By a simple comedy film.*
Negative	*- spy action flick with Antonio Banderas and Lucy Liu never comes together .*	*" It just felt... too heavy... for me to have it in my trunk."*	*- Espionage action with Antonio Banderas and Lucy Liu never comes together.*	*spy litigate flick with antonio banderas and lucy liu never comes together*	*The, is a fun and humorous story to write and read.*
	If you value your time and money , find an escape clause and avoid seeing this trite , predictable rehash .	*Isn't the best spy film ever made.*	*If you value your time and money, find an exit clause and avoid this mundane, predictable repetition.*	*if you value your time and money find an escape clause rehash avoid seeing this trite predictable and*	*A very satisfying viewing from a rare and very quiet spot.*
	No amount of burning , blasting , stabbing , and shooting can hide a weak script .	*It is a masterpiece, brilliantly directed, and incredibly well done.*	*No amount of firing, blasting, stabbing and shooting can hide a weak script.*	*no amount of burning blasting stabbing and bourgeon can hide a weak script*	*The documentary is packed with twists and turns and shocking revelations.*
	Suffers from a lack of clarity and audacity that a subject as monstrous and pathetic as Dahmer demands .	*The film itself has a very poor panorama, however, and it ultimately becomes an immensely enjoyable portrait.*	*Suffers from a lack of clarity and audacity demanded by a subject as monstrous and pathetic as Dahmer.*	*suffers from a lack of clarity and audacity that a subject as monstrous and pathetic as dahmer demands*	*This is a fairly short film.*
	Their parents would do well to cram earplugs in their ears and put pillowcases over their heads for 87 minutes .	*They live in a world that is almost as romantic as that little girl "Tiny Nuts" could be.*	*Her parents would do well to put earplugs in her ear and put pillowcases over her head for 87 minutes.*	*their parents would case do well to cram earplugs in their ears and put pillowcases over their heads for minutes*	*One "It 's a great story that requires some good humor and a little imagination.*
	This Tuxedo ... should have been sent back to the tailor for some major alterations .	*This is a film that I was unable to stomach... and very much enjoyed... until the last 30 minutes... when I finally got close to the film... and it turned into a garbage dump.*	*This tuxedo... should have been sent back to the tailor for some major changes.*	*this tuxedo should have been sent back to the tailor for john major some major alterations*	*No a delightful example of a film with great character and a wonderful cast of heroes.*

Table 3: A qualitative analysis of the diversity in the Sentiment Analysis task involving SST-5.