AI Open 4 (2023) 19–32

Contents lists available at ScienceDirect

AI Open

journal homepage: www.keaipublishing.com/en/journals/ai-open

Language as a latent sequence: Deep latent variable models for semi-supervised paraphrase generation

Jialin Yu [a,b,*], Alexandra I. Cristea [a,**], Anoushka Harit [a], Zhongtian Sun [a], Olanrewaju Tahir Aduragba [a], Lei Shi [a], Noura Al Moubayed [a]

[a] Department of Computer Science, Durham University, Stockton Road, Durham, DH1 3LE, UK
[b] Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK

ARTICLE INFO

Keywords:
Deep latent variable models
Paraphrase generation
Semi-supervised learning
Natural language processing
Deep learning

ABSTRACT

This paper explores deep latent variable models for semi-supervised paraphrase generation, where the missing target pair for unlabelled data is modelled as a latent paraphrase sequence. We present a novel unsupervised model named *variational sequence auto-encoding reconstruction* (**VSAR**), which performs latent sequence inference given an observed text. To leverage information from text pairs, we additionally introduce a novel supervised model we call *dual directional learning* (**DDL**), which is designed to integrate with our proposed VSAR model. Combining VSAR with DDL (**DDL+VSAR**) enables us to conduct semi-supervised learning. Still, the combined model suffers from a cold-start problem. To further combat this issue, we propose an improved weight initialisation solution, leading to a novel two-stage training scheme we call *knowledge-reinforced-learning* (**KRL**). Our empirical evaluations suggest that the combined model yields competitive performance against the state-of-the-art supervised baselines on complete data. Furthermore, in scenarios where only a fraction of the labelled pairs are available, our combined model consistently outperforms the strong supervised model baseline (**DDL**) by a significant margin ($p < .05$; Wilcoxon test). Our code is publicly available at https://github.com/jialin-yu/latent-sequence-paraphrase.

1. Introduction

Paraphrase generation is an important Natural Language Processing (NLP) problem, useful in many NLP applications, such as question answering (Dong et al., 2017), information retrieval (Lee et al., 2006), information extraction (Yao and Van Durme, 2014) and summarisation (Liu et al., 2008). Natural language itself is complicated and may be expressed in various alternative surface forms of the same underlying semantic content (Miller, 2019; Hosking et al., 2022). Hence it is critically important to integrate the paraphrase generation model as a component in real-world NLP systems, to offer robust responses to end users' inputs. Traditional solutions to paraphrase generation are generally rule-based (Kauchak and Barzilay, 2006; Narayan et al., 2016), utilising lexical resources, such as WordNet (Miller, 1992), to find word replacements. The recent trend brings to fore neural network models (Kumar et al., 2020; Zhou and Bhat, 2021; Meng et al., 2021; Su et al., 2021), which are typically based on a sequence-to-sequence learning paradigm (Sutskever et al., 2014).

These models have achieved remarkable success for paraphrase generation, due to complex architectures and sophisticated conditioning mechanisms, *e.g. soft, hard and self-attention*. However, the advancement of such models is primarily based on the availability of large-scale labelled data pairs. Instead, *this paper explores semi-supervised learning scenarios, where only a fraction of the labels are available*. This semi-supervised learning setting is favourable and extremely useful for industry scenarios (Zhu, 2005; Van Engelen and Hoos, 2020), due to the effort in terms of time and money to obtain good quality human annotations. A semi-supervised learning model often consists of two components: an unsupervised learning model and a supervised learning model.

Thus, for the unsupervised learning part, we propose a novel deep generative model, motivated by the classic variational autoencoder (VAE) (Kingma and Welling, 2014; Rezende et al., 2014; Mnih and Gregor, 2014), with an additional structural assumption tailored towards modelling language sequences, named *variational sequence auto-encoding reconstruction* (**VSAR**). To further explain, traditional VAEs typically embed data representations in a fixed latent space, with the general purpose of dimensionality reduction (Hinton and Salakhutdinov, 2006). Here we consider instead a latent variable in the form

* Corresponding author at: Department of Computer Science, Durham University, Stockton Road, Durham, DH1 3LE, UK.
** Corresponding author.
E-mail addresses: jialin.yu@durham.ac.uk (J. Yu), alexandra.i.cristea@durham.ac.uk (A.I. Cristea).

https://doi.org/10.1016/j.aiopen.2023.05.001
Received 7 May 2023; Accepted 18 May 2023
Available online 26 May 2023

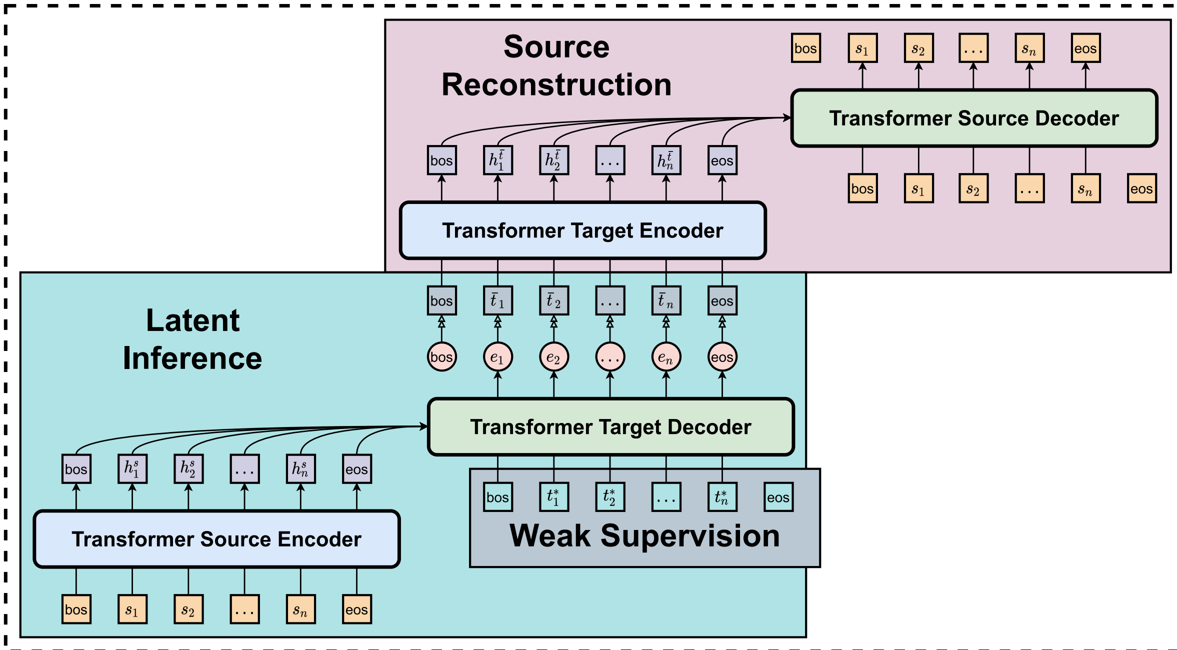

Fig. 1. Variational Sequence Auto-Encoding Reconstruction (**VSAR**).

of a discrete language sequence with various lengths. This assumption enforces more structural information to be adopted in the model training process and can additionally enhance the model interpretability, as language is naturally preserved as discrete variables (Fu et al., 2019). Following the recent prior works successfully incorporating discrete latent variables to improve paraphrasing (Hosking and Lapata, 2021; Hosking et al., 2022), we propose our model, the VSAR model, aimed to contain a more expressive form of discrete latent variable, shown in Fig. 1.

Furthermore, for the supervised learning part, motivated by dual learning (He et al., 2016; Su et al., 2019, 2020a,b), we propose a novel supervised model, named *dual directional learning* (**DDL**) that can be integrated with our proposed VAE model, which shares a part of the learning loop of the VSAR model. Combining both unsupervised and supervised models enables semi-supervised learning, by exploiting VAE's ability to marginalise latent variables for unlabelled data.

Our main original contributions in this paper thus include:

- presenting *the first study on semi-supervised learning for paraphrasing with deep discrete latent variable models*;
- introducing *two novel models: VSAR (unsupervised) and DDL (supervised)*, which can be combined for semi-supervised learning;
- proposing *a novel training scheme, knowledge-reinforced-learning (KRL)* to deal with the cold start problem in the combined semi-supervised model (DDL+VSAR);
- studying semi-supervised learning scenarios with the combined model on the full data and empirically showing that our model achieves competitive state-of-the-art results;
- presenting a study of semi-supervised scenarios on a fraction of the labelled data(i.e., when incorporating unlabelled data), demonstrating significantly better results for our models than for very strong supervised baselines.

2. Related work

2.1. Paraphrase generation

Paraphrases express the surface forms of the underlying semantic content (Hosking et al., 2022) and capture the essence of language diversity (Pavlick et al., 2015). Early work on automatic generation of paraphrases are generally rule-based (Kauchak and Barzilay, 2006; Narayan et al., 2016), but the recent trend brings to the fore neural network solutions (Gupta et al., 2018; Fu et al., 2019; Kumar et al., 2020; Meng et al., 2021; Su et al., 2021; Hosking and Lapata, 2021; Hosking et al., 2022; Chen et al., 2022; Xie et al., 2023). Current research for paraphrasing mainly focuses on supervised methods, which require the availability of a large number of source and target pairs. In this work, we instead explore a semi-supervised paraphrasing method, where only a fraction of source and target pairs are observed, and where a large number of unlabelled source texts exist. We made an assumption that each missing target text can be considered as a latent variable in deep generative models. Thus, for unsupervised data, each missing paraphrase output is modelled as a latent variable. Compared to the standard approach, where the semantics of a sentence is presented as a dense high-dimensional vector, this assumption enforces the model to learn more *structured* representations. Hence, our proposed model can be considered as a type of deep latent structure model (Martins et al., 2019). In this paper, we present two models and combine them for paraphrasing: one for unsupervised learning and one for supervised learning. Our combined model extends the idea in Miao and Blunsom (2016), Fu et al. (2019); compared with Fu et al. (2019), our model utilises conversely a more natural language structure (an ordered sequence other than an unordered bag of words); compared with Miao and Blunsom (2016), our model utilises a self-attention mechanism other than convolution operations and has the benefit of being able to model various lengths of latent sequences, rather than a fixed length.

Furthermore, our combined model is associated with prior works that introduce a discrete latent variable (Hosking and Lapata, 2021; Hosking et al., 2022), and it uses an arguably more expressive latent variable, in the form of language outputs.

2.2. Deep latent variable models for text

Deep latent variable models have been studied for text modelling (Miao et al., 2016; Kim et al., 2018). The most common and widely adopted latent variable model is the standard VAE model with a Gaussian prior (Bowman et al., 2016), which suffers from posterior collapse (Dieng et al., 2019; He et al., 2019). Multiple studies have been conducted to combat this issue (Higgins et al., 2017; Razavi et al., 2019; Wang et al., 2021). In particular, β-VAE (Higgins et al., 2017) introduces a penalty term to balance VAE reconstruction and prior regularisation intuitively and is adopted as one of our baselines.

While much of the research focuses on continuous latent variable models (Miao et al., 2016; Kim et al., 2018), the text is naturally presented in discrete form and may not be well represented with continuous latent variables. Early work on discrete deep latent variable models (Miao and Blunsom, 2016; Wen et al., 2017) adopted the REINFORCE algorithm (Mnih and Gregor, 2014; Mnih et al., 2014); however, it suffers from very high variance. With the recent advancement in statistical relaxation techniques, the Gumbel trick (Jang et al., 2017; Maddison et al., 2017) was utilised, to model discrete structures in the latent variable model of the text (Choi et al., 2018; Fu et al., 2019; He et al., 2020; Mercatali and Freitas, 2021). Our work adopts the Gumbel-trick with subset sampling (Xie and Ermon, 2019) for natural language generation tasks and, *for the first time, studies discrete language sequences as a latent variable for the paraphrasing task*. Our proposed model is strongly associated with Miao and Blunsom (2016), He et al. (2020); however, we study the problem under the semi-supervised setup for the paraphrase generation tasks. Furthermore, we present a novel inference algorithm (our knowledge-reinforced-learning (KRL) scheme) to help aid learning in deep generative models and achieve competitive performance for both full data and incomplete (fractional) data settings. In terms of deep latent variable models for text, there has been a recent surge of interest in learning discrete latent structures (Niculae et al., 2023). In this paper, we contribute thus to this research field by modelling the latent variable as a language sequence. Additionally, some recent work focused on combining latent variable models with diffusion models, to achieve state-of-the-art performance (Huang et al., 2023; Schneider et al., 2023) for text-to-audio generation tasks.

3. Variational Sequence Auto-Encoding Reconstruction (VSAR)

In this section, we present the VSAR model (Fig. 1).[1] The model consists of four separate neural network models - a source encoder, a target decoder, a target encoder, and a source decoder. Under the unsupervised learning setup, we only observe source text $\boldsymbol{s}$ and no target text $\boldsymbol{t}$. We reformulate the problem of modelling fully observed source text $\boldsymbol{s}$, as modelling the partially observed parallel source text $\boldsymbol{s}$ and its associated latent target pair $\bar{\boldsymbol{t}}$. We adopt Bayesian inference, to marginalise the latent target string $\bar{\boldsymbol{t}}$ from the joint probability distribution $p_\theta(s, \bar{t})$, based on Eq. (7), as shown in Fig. 1.

Thus, in the VSAR model, the **latent inference** network, parameterised as $q_\phi(\bar{t}|s)$, takes source text $\boldsymbol{s}$ and generates a latent target sample $\bar{\boldsymbol{t}}$. The **source reconstruction** network, parameterised as $p_\theta(s|\bar{t})$, reconstructs the observed source text $\boldsymbol{s}$ back, based on the latent target sample $\bar{\boldsymbol{t}}$. As the prior distribution, a language model is pre-trained on unlabelled source text corpus, to approximate the prior distribution $p(\bar{t})$.[2] The prior is introduced for regularisation purposes (Miao and Blunsom, 2016; He et al., 2020), which enforces that samples are more likely to be 'reasonable' natural language text.

Motivated by the benefits of **parameters sharing** in multi-task learning for natural language generation (Luong et al., 2016; Guo et al., 2018a,b; Wang et al., 2020), we share model parameters for the source encoder and the target encoder, denoted as $\boldsymbol{f}_{encode}$; similarly, we share model parameters for the source decoder and the target decoder, denoted as $\boldsymbol{f}_{decode}$. In the following sections, we use $\boldsymbol{f}_{encode}$ and $\boldsymbol{f}_{decode}$ to represent all encoders and decoders in the VSAR model, respectively.

[1] The language model prior and weak supervision decoding is omitted, for clarity.

[2] We leverage linguistic knowledge of the paraphrase generation task, in which a paraphrase text string can be considered as its own paraphrase.

3.1. Weak supervision

In the VSAR model, we empirically found that the quality of the latent sequence $\bar{\boldsymbol{t}}$ is very unstable, especially at the beginning of the training. To combat this issue, motivated by the idea of weak supervision (Du et al., 2021; Chang et al., 2021), we propose to use pseudo-labels to guide VSAR throughout training. Before each model performs the forward-pass using the back-propagation algorithm, we first assign pseudo-labels to each token in the unobserved latent target sample $\bar{\boldsymbol{t}}$ based on the current model parameter (from the previous iteration). The pseudo-labels are detached from the computational graph; hence no gradient is updated during the weak supervision process. The pseudo-labels can be considered as a weak supervision signal for 'teacher forcing training' (Williams and Zipser, 1989).

The encoder model takes the source string $\boldsymbol{s} = (s_1, \ldots, s_n)$ as input and produces its corresponding contextual vector $\boldsymbol{h}^s = (h_1^s, \ldots, h_n^s)$:

$$\boldsymbol{h}^s = \boldsymbol{f}_{encode}(\boldsymbol{s}) \tag{1}$$

We adopt a greedy decoding scheme to assign pseudo-target labels $\boldsymbol{t}^*$ and assume that a good paraphrase ought to have a similar length as the original sentence (Burrows et al., 2013; Cao et al., 2017); such that $\boldsymbol{t}^* = (t_1^*, \ldots, t_n^*)$. Let t_i^* be the ith word in the pseudo target sequence; we construct this sequence in an auto-regressive manner:

$$t_i^* = \boldsymbol{f}_{decode}(\boldsymbol{h}^s; t_{1:i-1}^*) \tag{2}$$

3.2. Target inference

Once the pseudo-target labels $\boldsymbol{t}^*$ are assigned, we perform latent variable inference with the latent inference network. Since the source string $\boldsymbol{s}$ remains the same, we reuse the value of the contextual vector $\boldsymbol{h}^s$ in the weak supervision section. Let $\bar{t}_j$ be the jth words in the latent sample and e_j be the corresponding output of the target decoder model. We construct the latent sample $\bar{\boldsymbol{t}}$ using contextual vector $\boldsymbol{h}_s$ and all $t_{1:j-1}^*$ words in the pseudo-labels:

$$\begin{aligned} e_j &= \boldsymbol{f}_{decode}(\boldsymbol{h}_s; t_{1:j-1}^*) \\ \bar{t}_j &\sim Gumbel\text{-TOP}k(e_j, \tau) \end{aligned} \tag{3}$$

Here, $\bar{t}_i$ is drawn via the Gumbel trick (Jang et al., 2017; Maddison et al., 2017) with temperature τ as an additional hyper-parameter, which controls the probability distribution of the samples. At a high temperature τ, we equivalently sample from a uniform distribution; at a low temperature τ, we equivalently sample from a categorical distribution. Due to the enormous size of the vocabulary, sampling from this simplex can be difficult, hence we further adopt the TOP-k subset sampling technique (Xie and Ermon, 2019) to improve sampling efficiency.

We explore two different schemes commonly used in the literature: (1) we use a fixed temperature τ of 0.1, as in Chen et al. (2018); and (2) we gradually anneal the temperature τ from a high temperature of 10 to a low temperature of 0.01, as in Balin et al. (2019). Our empirical results suggest that annealing the temperature τ during training yields significantly better results ($p < .05$; Wilcoxon test), which are thus used to report the final results. We use a k-value of 10, as suggested in Fu et al. (2019).

3.3. Source reconstruction

For the source reconstruction network, the encoder model takes the latent target sequence string $\bar{\boldsymbol{t}} = (\bar{t}_1, \ldots, \bar{t}_n)$ as input and produces its corresponding contextual vector $\boldsymbol{h}^{\bar{t}} = (h_1^{\bar{t}}, \ldots, h_n^{\bar{t}})$:

$$\boldsymbol{h}^{\bar{t}} = \boldsymbol{f}_{encode}(\bar{\boldsymbol{t}}) \tag{4}$$

Let $\hat{s}_k$ be the kth word in the reconstructed source string, during the training; we retrieve the reconstructed source string $\hat{s}$ via:

$$\hat{s}_k = \boldsymbol{f}_{decode}(\boldsymbol{h}^{\bar{t}}; s_{1:k-1}) \tag{5}$$

Table 1
Semi-supervised learning experiment results for quora.

Model	Labelled	Unlabelled	B-1	B-2	B-3	B-4	i-B	R-1	R-2	R-L
DDL	20K	–	46.68	33.44	25.46	20.18	11.08	47.57	25.42	45.50
DDL+VSAR$_1$	20K	20K	47.80 ↑	34.33 ↑	26.17 ↑	20.76 ↑	11.25 ↑	48.03 ↑	25.82 ↑	45.84 ↑
DDL+VSAR$_2$	20K	100K	**50.26** ↑	**36.87** ↑	**28.50** ↑	**22.82** ↑	**11.60** ↑	**51.51** ↑	**28.45** ↑	**49.07** ↑
DDL	50K	–	53.31	40.22	31.70	25.80	**13.80**	**55.63**	**32.15**	**53.13**
DDL+VSAR$_1$	50K	50K	53.33 ↑	39.93 ↓	31.39 ↓	25.49 ↓	13.45 ↓	55.51 ↓	31.90 ↓	52.95 ↓
DDL+VSAR$_2$	50K	100K	**53.79** ↑	**40.47** ↑	**31.86** ↑	**25.93** ↑	13.67 ↓	55.58 ↓	31.89 ↓	52.93 ↓

Table 2
Semi-supervised learning experiment results for MSCOCO.

Model	Labelled	Unlabelled	B-1	B-2	B-3	B-4	i-B	R-1	R-2	R-L
DDL	20K	–	66.82	47.25	33.14	23.75	16.66	40.53	14.95	36.94
DDL+VSAR$_1$	20K	20K	66.98 ↑	47.28 ↑	33.10 ↓	23.72 ↓	16.54 ↓	40.60 ↑	14.95 ↑	36.94 ↑
DDL+VSAR$_2$	20K	93K	**67.64** ↑	**48.00** ↑	**33.96** ↑	**24.55** ↑	**16.68** ↑	**40.87** ↑	**15.12** ↑	**37.01** ↑
DDL	50K	–	69.39	50.17	36.06	26.49	18.43	42.08	16.31	38.27
DDL+VSAR$_1$	50K	50K	69.43 ↑	50.21 ↑	36.08 ↑	26.45 ↓	18.31 ↓	42.20 ↑	16.33 ↑	38.31 ↑
DDL+VSAR$_2$	50K	93K	**69.91** ↑	**50.65** ↑	**36.52** ↑	**26.93** ↑	**18.51** ↑	**42.39** ↑	**16.46** ↑	**38.40** ↑

3.4. Learning and inference for VSAR

In the SVAR model, there are two sets of parameters, $\boldsymbol{\phi}$ and θ, which are required to be updated. Let $\boldsymbol{S}$ be the observed random variable for the source text, $\bar{\boldsymbol{T}}$ be the latent random variable for the target text, and N be the total number of the unlabelled source text. We have the following joint likelihood for the SVAR model, parameterised by θ:

$$p(\boldsymbol{S}, \bar{\boldsymbol{T}}; \theta) = \prod_{i=1}^{N} p(s_{(i)}|\bar{t}_{(i)}; \theta)p(\bar{t}_{(i)}) \tag{6}$$

The log marginal likelihood $\boldsymbol{L}_1$ of the observed data that will be approximated during training is $\log p(\boldsymbol{S}; \theta)$. We adopt amortised variational inference (Kingma and Welling, 2014; Rezende et al., 2014; Mnih and Gregor, 2014) and build a surrogate function, approximated with a neural network $q(\bar{\boldsymbol{T}}|\boldsymbol{S}; \boldsymbol{\phi})$, parameterised by $\boldsymbol{\phi}$, to derive the evidence lower bound (ELBO) for the joint likelihood:

$$\begin{aligned} \boldsymbol{L}_1 &= \log \sum_{\bar{\boldsymbol{T}}} p(\boldsymbol{S}, \bar{\boldsymbol{T}}; \theta) \geq \mathcal{L}_{ELBO}(\boldsymbol{S}, \bar{\boldsymbol{T}}; \theta, \boldsymbol{\phi}) \\ &= \sum_{i=1}^{N} \{\mathbb{E}_{q(\bar{t}|s_{(i)};\boldsymbol{\phi})}[\log p(s_{(i)}|\bar{t}; \theta)] - \mathbb{D}_{KL}[q(\bar{t}|s_{(i)}; \boldsymbol{\phi})||p(\bar{t})]\} \end{aligned} \tag{7}$$

The most common variational family in the VAE framework relies on the reparameterisation trick (Kingma and Welling, 2014), which is not applicable to the non-differentiable discrete latent variable. An approach for optimising learning with such latent variables uses the REINFORCE algorithm (Mnih and Gregor, 2014; Mnih et al., 2014); however, this algorithm generally suffers from high variance. In this paper, we instead use Gumbel-Softmax (Jang et al., 2017; Maddison et al., 2017) with differentiable subset sampling (Xie and Ermon, 2019), to retrieve top-k samples without replacement. Nevertheless, since sampling a one-hot form vector induces high variance, we apply the straight-through technique (Bengio et al., 2013) as a biased estimator of the gradient, to combat this variance.

During training, while optimising the log-likelihood, we perform learning (θ) and inference ($\boldsymbol{\phi}$) at the same time. The parameters are jointly optimised with the same optimiser. Since we are sharing parameters in our model, in practice, we are updating the same set of parameters (shared by θ and $\boldsymbol{\phi}$) with source data only.

4. Dual Directional Learning (DDL)

In this section, we introduce the Dual Directional Learning (DDL) model, which we use for supervised paraphrase generation. The DDL model consists of two sets of standard Transformer models (Vaswani et al., 2017), each with its own two separate neural networks — an encoder and a decoder. We perform standard sequence-to-sequence learning, with the fully observed parallel source text $\boldsymbol{s}$ and its associated target pair $\boldsymbol{t}$, in dual directions. The **target generation** network $p_{\theta_{t|s}}(\boldsymbol{t}|\boldsymbol{s})$ takes source text $\boldsymbol{s}$ as input and generates target text $\boldsymbol{t}$; and the **source generation** network $p_{\theta_{s|t}}(\boldsymbol{s}|\boldsymbol{t})$ takes target text $\boldsymbol{t}$ as input and generates source text $\boldsymbol{s}$.

4.1. Parameter learning

In the DDL model, there are two sets of parameters, $\boldsymbol{\theta}_{s|t}$ and $\boldsymbol{\theta}_{t|s}$, which are required to be updated. Let $\boldsymbol{S}$ be the observed random variable for source text, $\boldsymbol{T}$ be the observed random variable for target text, and M be the number of labelled pairs; we then have the following conditional likelihood for our DDL model:

$$\begin{aligned} p(\boldsymbol{S}|\boldsymbol{T}; \boldsymbol{\theta}_{s|t}) &= \prod_{i=1}^{M} p(s_{(i)}|t_{(i)}; \boldsymbol{\theta}_{s|t}) \\ p(\boldsymbol{T}|\boldsymbol{S}; \boldsymbol{\theta}_{t|s}) &= \prod_{i=1}^{M} p(t_{(i)}|s_{(i)}; \boldsymbol{\theta}_{t|s}) \end{aligned} \tag{8}$$

The log conditional likelihood $\boldsymbol{L}_2$ of the observed data pairs can be jointly learnt during training as:

$$\boldsymbol{L}_2 = \sum_{i=1}^{M} (\log p(s_{(i)}|t_{(i)}; \boldsymbol{\theta}_{s|t}) + \log p(t_{(i)}|s_{(i)}; \boldsymbol{\theta}_{t|s})) \tag{9}$$

During training, we perform dual learning ($\boldsymbol{\theta}_{s|t}$ and $\boldsymbol{\theta}_{t|s}$) at the same time and the parameters are jointly optimised with the same optimiser.

4.2. Parameter sharing

Once again, motivated by the benefits of multi-task learning for natural language generation (Luong et al., 2016; Guo et al., 2018a,b; Wang et al., 2020), we share model parameters for the target generation and the source generation network. Although sharing parameters is a very simple technique, as shown in Tables 1 and 2, the DDL model significantly improves the performance of paraphrase generation with respect to the Transformer baseline ($p < .05$; Wilcoxon test), which only handles sequence-to-sequence learning in a single direction.

5. Combining VSAR and DDL for semi-supervised learning

In this section, we introduce our semi-supervised learning model (VSAR+DDL), which combines models presented in previous sections.

For semi-supervised learning, the log-likelihood of the data can be expressed as follow:

$$
\begin{aligned}
\boldsymbol{L} &= \boldsymbol{L}_1 + \boldsymbol{L}_2 \\
&= \sum_{i=1}^{N} \{\mathbb{E}_{q(\bar{t}|s_{(i)};\boldsymbol{\phi})}[\log p(s_{(i)}|\bar{t};\boldsymbol{\theta})] - \mathbb{D}_{KL}[q(\bar{t}|s_{(i)};\boldsymbol{\phi})||p(\bar{t})]\} \\
&+ \sum_{i=1}^{M} (\log p(s_{(i)}|t_{(i)};\boldsymbol{\theta}_{s|t}) + \log p(t_{(i)}|s_{(i)};\boldsymbol{\theta}_{t|s}))
\end{aligned}
\tag{10}
$$

As suggested in Eq. (10), for unsupervised learning and supervised learning, the likelihood function involves the same set of conditional probability between s and t. We hypothesise that sharing parameters between these two models is beneficial. Thus, we share two sets of neural network parameters from the VSAR and DDL models (i.e. $q_{\boldsymbol{\phi}}(\bar{t}|s) \equiv p_{\boldsymbol{\theta}_{t|s}}(t|s)$ and $p_{\boldsymbol{\theta}}(s|\bar{t}) \equiv p_{\boldsymbol{\theta}_{s|t}}(s|t)$). This allows the strong supervision signal from the DDL model to contribute to the VSAR model, directly. At the same time, the unsupervised signal from the VSAR model can benefit the generalisation of the DDL model.

5.1. Knowledge reinforced learning

Our empirical experiments suggest that our combined model (DDL+VSAR) suffers from a cold-start problem for parameter optimisation, when conducting semi-supervised learning from scratch. We found that a key to the success of our model is to have **better initialisation** of the model weight. Hence, we present a novel training scheme called **knowledge reinforced learning** (Fig. 2), which includes two-stage training. In stage one (pre-training), we conduct supervised learning with our DDL model on paired training sets, as demonstrated in Algorithm 1. In stage two (fine-tuning), we initialise the VSAR model parameter with the best performance DDL model from stage one; and we conduct semi-supervised learning with labelled and unlabelled data, as demonstrated in Algorithm 2. The intuition is to inject better preliminary information into training the VSAR model.

Algorithm 1 Knowledge Reinforced Pre-Training

Input:
Supervised Training Data ($\mathcal{D}_T^S = \{(s_1, t_1), ..., (s_N, t_N)\}$), Supervised Validation Data ($\mathcal{D}_V^S$)
Parameter:
DDL Model: $\boldsymbol{\theta}_{s|t}$ and $\boldsymbol{\theta}_{t|s}$
Parameter Sharing:
Set $\boldsymbol{\theta}_{s|t}$ equals to $\boldsymbol{\theta}_{t|s}$ throughout knowledge reinforced pre-training
Output: $\boldsymbol{\theta}_{s|t}{}^*$ and $\boldsymbol{\theta}_{t|s}{}^*$

```
Initialise θ_{s|t} and θ_{t|s} with a random seed; set maximum training epochs as T; set L_2* = 0
while maximum epochs not reached do
    Update θ_{s|t} and θ_{t|s} with mini-batch data from D_T^S based on Equation (9)
    if L_2 in Equation (9) calculated based on D_V^S larger than L_2* then
        Set L_2* ← L_2
        Set θ_{s|t}* ← θ_{s|t}
        Set θ_{t|s}* ← θ_{t|s}
    end if
end while
Return: θ_{s|t}* and θ_{t|s}*
```

5.2. Effect of language model prior

In literature (Higgins et al., 2017; Miao and Blunsom, 2016; Yang et al., 2018; He et al., 2020), a language model prior is introduced for regularisation purposes, which enforces samples to more likely contain

Algorithm 2 Knowledge Reinforced Fine-Training

Input:
Unsupervised Data ($\mathcal{D}^U = \{s_1, ..., s_M\}$)
Supervised Training Data ($\mathcal{D}_T^S = \{(s_1, t_1), ..., (s_N, t_N)\}$), Supervised Validation Data ($\mathcal{D}_V^S$)
Parameter:
VSAR Model: $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$; DDL Model: $\boldsymbol{\theta}_{s|t}$ and $\boldsymbol{\theta}_{t|s}$
Parameter Sharing:
Set $\boldsymbol{\phi}$ equals to $\boldsymbol{\theta}_{t|s}$; $\boldsymbol{\theta}$ equals to $\boldsymbol{\theta}_{s|t}$; and $\boldsymbol{\theta}_{s|t}$ equals to $\boldsymbol{\theta}_{t|s}$ throughout knowledge reinforced fine-tuning
Output: $\boldsymbol{\theta}_{s|t}{}^{**}$, $\boldsymbol{\theta}_{t|s}{}^{**}$; $\boldsymbol{\phi}^{**}$ and $\boldsymbol{\theta}^{**}$

```
Initialise φ and θ_{t|s} with θ_{t|s}*; and initialise θ and θ_{s|t} with θ_{s|t}*; set maximum training epochs as T; set L_2* = 0.
while maximum epochs not reached do
    Update θ_{s|t} and θ_{t|s} with mini-batch data from D_T^S based on Equation (9)
    Update φ and θ with mini-batch data from D^U based on Equation (7)
    if L_2 in Equation (9) calculated based on D_V^S larger than L_2* then
        Set L_2* ← L_2
        Set θ_{s|t}** ← θ_{s|t}
        Set θ_{t|s}** ← θ_{t|s}
        Set φ** ← φ
        Set θ** ← θ
    end if
end while
Return: θ_{s|t}**, θ_{t|s}**; φ** and θ**
```

a 'reasonable' natural language, especially at the beginning of the training. Hence, we adopt the same approach and use a prior in our model. We empirically found the prior useful when the labelled dataset was relatively small. However, surprisingly, we found that training without a prior in the VSAR model yields better results with our parameter initialisation method, when the dataset is large. The improvement is significant ($p < .05$; Wilcoxon test), as shown in Tables 5 and 6. We report the results without language model prior as DDL +VSAR*, and the log-likelihood becomes:

$$
\begin{aligned}
\boldsymbol{L}^* &= \sum_{i=1}^{N} \{\mathbb{E}_{q(\bar{t}|s_{(i)};\boldsymbol{\phi})}[\log p(s_{(i)}|\bar{t};\boldsymbol{\theta})]\} \\
&+ \sum_{i=1}^{M} (\log p(s_{(i)}|t_{(i)};\boldsymbol{\theta}_{s|t}) + \log p(t_{(i)}|s_{(i)};\boldsymbol{\theta}_{t|s}))
\end{aligned}
\tag{11}
$$

To further investigate this issue, we conducted experiments to compare the performance of semi-supervised learning when training with prior (Eq. (10)) or without prior (Eq. (11)) under different data portion setting. We empirically found that with a low portion of labelled data, the combined model (DDL+VSAR) with a prior grants significantly ($p < .05$; Wilcoxon test) better performance and is more stable. This aligns with the observations in Higgins et al. (2017), Miao and Blunsom (2016), Yang et al. (2018), He et al. (2020). However, with a large portion of labelled data, the combined model (DDL+VSAR) without the prior is significantly ($p < .05$; Wilcoxon test) better.

We argue that this phenomenon relates to our choice of prior, as it is pre-trained on an unlabelled source text corpus, instead of on the target text corpus. This approximation leads to a distribution shift from the true prior distribution $p(\bar{t})$. Thus, when a low portion of the labelled data is used in Algorithm 1, the final DDL parameters $\theta^*_{s|t}$ and $\theta^*_{t|s}$ for the initialisation VSAR model in Algorithm 2 are not good enough. The prior, in this case, can still benefit the combined model in the semi-supervised learning setting. However, with a large portion of labelled data, the initialisation is good enough, and, in such a case, the distribution shift can harm the combined model.

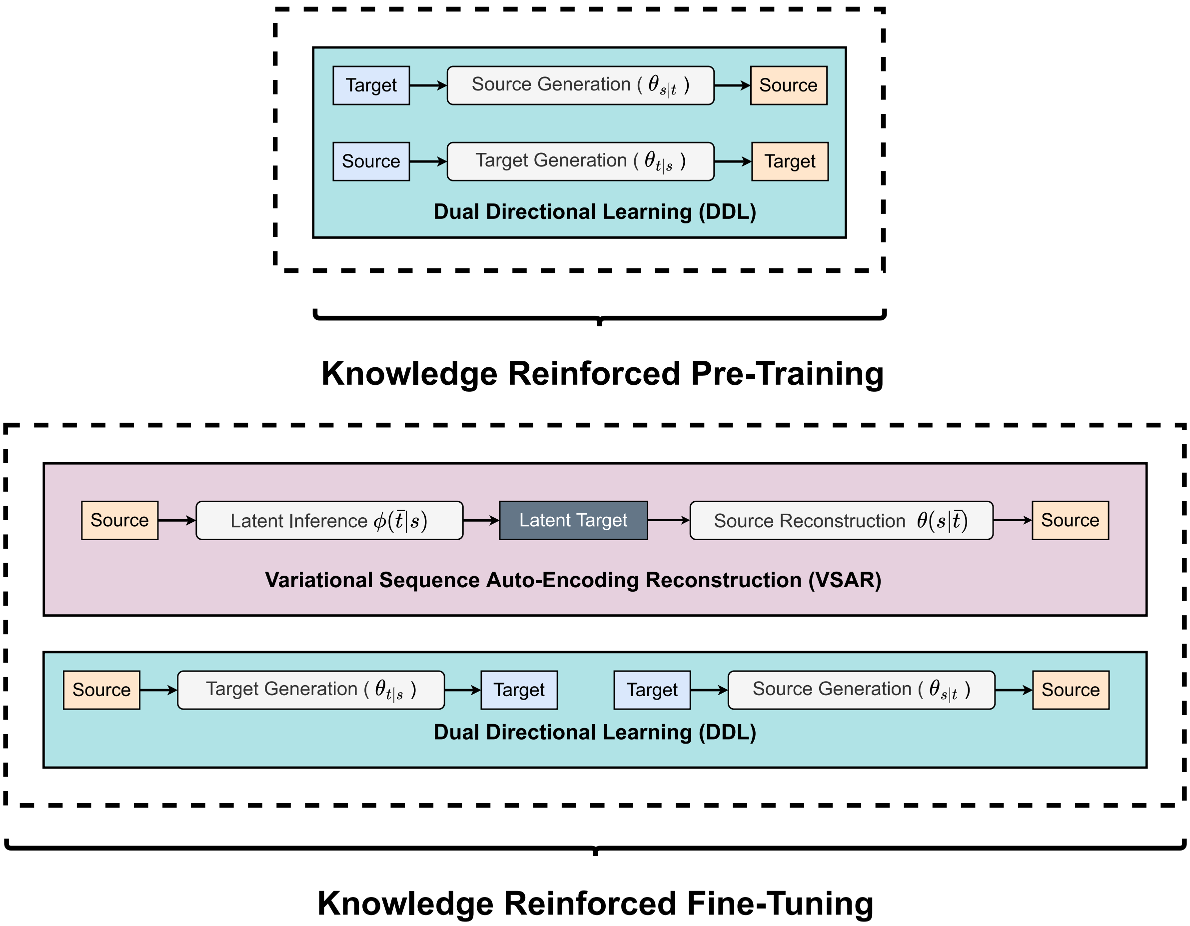

Fig. 2. Knowledge Reinforced Learning (**KRL**).

5.3. Semi-supervised learning setup

Under the semi-supervised learning setting, we limit the size of the supervised source and target pairs to be less than or equal to the unsupervised source text ($M \leq N$), as we could otherwise just conduct supervised learning, to take full advantage of the observed data pairs. This paper presents a thorough study of different sizes for M and N. Experimental results under this setting are presented in Table 1, Table 2, Tables 3 and 4.

6. Experiments

Here, we describe the datasets, experimental setup, evaluation metrics and experimental results.

6.1. Datasets

MSCOCO (Lin et al., 2014): This dataset has been widely adopted to evaluate paraphrase generation methods and contains human-annotated captions of images. Each image is associated with five captions from different annotators, who describe the most prominent object or action in an image. We use the 2017 version for our experiments; from the five captions accompanying each image, we randomly choose one as the source string and one as the target string for training. We randomly choose one as the source string for testing and use the rest four as the references.

Quora[3]: This dataset consists of 150K lines of question duplicate pairs, and it has been used as a benchmark dataset for paraphrase generation since 2017. However, since this dataset does not contain a specific split for training and testing, prior models are evaluated based on different subset sizes of data.

For both datasets (MSCOCO and Quora), in order to improve the reproducibility of our results, we use a pre-trained tokeniser ('bert-base-uncased' version) from Devlin et al. (2019)[4] and set the maximum token length as 20 (by removing the tokens beyond the first 20). Following Li et al. (2019), Fu et al. (2019), Su et al. (2021), we use training, validation and test sets as 100K, 4K and 20K, respectively for the Quora dataset; and 93K, 4K and 20K, respectively, for MSCOCO. For the complementary study in Tables 7 and 8, we use training, validation and test sets as 100K, 24K and 24K for the Quora dataset; and 100K, 5K and 5K for MSCOCO, in order to have a fair comparison with the results reported in Hosking and Lapata (2021), Hosking et al. (2022).

Other available datasets for paraphrase generation tasks include: ParaBank (Hu et al., 2019) and PARANMT (Wieting and Gimpel, 2017), which are two large-scale datasets created using back translation techniques from another non-English parallel corpus. Since these two datasets are less adopted by researchers in the literature, we cannot directly compare them against existing works. Still, as an alternative, to further demonstrate the efficacy of our proposed models, we conduct semi-supervised learning experiments on these two datasets. For the ParaBank dataset, we took the 'ParaBank v1.0 (5 m pairs)'[5] and use a similar setup as for the Quora dataset. The dataset consists of 5M lines of duplicated pairs; we randomly choose the same subset of the data for our experiments and use training, validation and test set of size 100K,

3 https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

4 https://github.com/huggingface/transformers

5 https://nlp.jhu.edu/parabank/

Table 3
Semi-supervised learning experiment results for ParaBank.

Model	Labelled	Unlabelled	B-1	B-2	B-3	B-4	i-B	R-1	R-2	R-L
DDL	20K	–	50.48	39.21	31.11	25.16	12.98	57.03	33.77	55.31
DDL+VSAR$_1$	20K	20K	50.69 ↑	39.41 ↑	31.35 ↑	25.45 ↑	13.14 ↑	56.99 ↓	33.92 ↑	55.20 ↓
DDL+VSAR$_2$	20K	100K	**54.04** ↑	**43.41** ↑	**35.44** ↑	**29.40** ↑	**14.94** ↑	**60.66** ↑	**37.50** ↑	**58.90** ↑
DDL	50K	–	58.05	48.84	41.59	35.82	18.19	66.45	44.63	64.81
DDL+VSAR$_1$	50K	50K	58.29 ↑	49.24 ↑	42.10 ↑	36.40 ↑	18.45 ↑	66.88 ↑	45.23 ↑	65.23 ↑
DDL+VSAR$_2$	50K	100K	**59.22** ↑	**50.33** ↑	**43.23** ↑	**37.51** ↑	**18.89** ↑	**67.64** ↑	**46.05** ↑	**65.98** ↑

Table 4
Semi-supervised learning experiment results for PARANMT.

Model	Labelled	Unlabelled	B-1	B-2	B-3	B-4	i-B	R-1	R-2	R-L
DDL	20K	–	62.49	49.47	39.92	32.84	17.40	63.13	40.43	61.15
DDL+VSAR$_1$	20K	20K	63.48 ↑	50.70 ↑	41.32 ↑	34.33 ↑	18.16 ↑	64.59 ↑	41.97 ↑	62.58 ↑
DDL+VSAR$_2$	20K	100K	**66.46** ↑	**54.20** ↑	**44.80** ↑	**37.60** ↑	**19.42** ↑	**67.55** ↑	**45.25** ↑	**65.49** ↑
DDL	50K	–	70.55	59.53	50.79	43.78	22.82	71.97	51.30	69.97
DDL+VSAR$_1$	50K	50K	70.41 ↓	59.48 ↓	50.81 ↑	43.87 ↑	23.02 ↑	72.06 ↑	51.45 ↑	70.07 ↑
DDL+VSAR$_2$	50K	100K	**70.91** ↑	**60.12** ↑	**51.48** ↑	**44.53** ↑	**23.25** ↑	**72.57** ↑	**52.03** ↑	**70.57** ↑

Table 5
Experiment results for Quora.

Model	B-1	B-2	B-3	B-4	i-B	R-1	R-2	R-L
Upper Bound (Copy Source)	63.36	49.99	40.47	33.54	–	63.04	38.15	59.64
Lower Bound (Random Select)	16.10	4.50	1.94	0.79	–	9.13	1.54	8.79
Residual-LSTM (Prakash et al., 2016)	53.59	39.49	30.25	23.69	15.93	55.10	33.86	53.61
β-VAE (Higgins et al., 2017)	47.86	33.21	24.96	19.73	10.28	47.62	25.49	45.46
Transformer (Vaswani et al., 2017)	53.56	40.47	32.11	25.01	17.98	57.82	32.58	56.26
LBOW-TOPk (Fu et al., 2019)	55.79	42.03	32.71	26.17	19.03	58.79	34.57	56.43
IANet+X (Su et al., 2021)	56.06	42.69	33.38	26.52	19.62	59.33	35.01	57.13
Transformer (our implementation)	54.73	41.59	32.96	26.94	14.50	56.90	33.28	54.29
DDL (our model)	55.97 ↑	43.02 ↑	34.32 ↑	28.19 ↑	14.83 ↑	**58.80** ↑	35.00 ↑	56.11 ↑
DDL + SVAR (our model)	55.79 ↑	42.79 ↑	34.11 ↑	28.01 ↑	**14.92** ↑	58.61 ↑	34.75 ↑	55.91 ↑
DDL + SVAR* (our model)	**55.99** ↑	**43.05** ↑	**34.37** ↑	**28.23** ↑	14.81 ↑	58.79 ↑	**35.02** ↑	**56.14** ↑

4K and 20K, respectively. For the PARANMT dataset, we first filter out the paraphrase pairs (remove entries where the paragram-phrase score[6] is higher than 0.95 and smaller than 0.90; set the maximum token length as 20 and the minimum token length as 5) and keep the middle percentiles, as recommended in Wieting and Gimpel (2017), to remove noisy and trivial paraphrases. After the filtering, we apply a similar experimental setup as the Quora dataset and randomly choose the same subset of the data for our experiments and use training, validation and test set of size 100K, 4K and 20K, respectively.

6.2. Baselines

We consider several state-of-the-art baselines, presented in Table 5, Table 6, Table 7, and Table 8. Note that these experimental results are directly taken from Su et al. (2021)[7] and Hosking et al. (2022). For evaluation, we start with our implementation of the Transformer model as the absolute baseline, which achieves competitive performance, as reported in Su et al. (2021). The Transformer model (Vaswani et al., 2017) is considered as the SOTA model, which is very ‘hard to beat’. We report our model performance based on a similar setup as in Su et al. (2021) and Hosking et al. (2022).

Recently, large-scale pre-trained language models (PLMs) have been widely adopted as the state-of-the-art approaches for both understanding and generation tasks in the NLP domain; in this paper, however, PLMs are not selected as the baseline model to compete against, as they contain external information trained in an unsupervised fashion based on a large-scale text corpus. Alternatively, in this paper, we focused on end-to-end learning for paraphrase generation tasks from scratch. This experimental setting allows us to directly compare with other literature. Hence, in this paper, we are not compared against any PLMs, other than leaving this as future work. However, note that all types of PLM models are primarily based on the Transformer architecture, which we used as a baseline model to compare with. The implication is that our methods could potentially be used for improving the performance of PLMs with limited labelled resources and large-scale unlabelled data points.

6.3. Experimental setup

In this section, we introduce our primary experimental setup. We do not use any external word embedding, such as Glove (Pennington et al., 2014), word2vec (Mikolov et al., 2013) or BERT (Devlin et al., 2019) for initialisation; rather, we obtain word embedding with end-to-end training, in order not to use any prior knowledge and better understand the impact of our model. We use the ‘base’ version of the Transformer model (Vaswani et al., 2017), which is a 6-layer model with 512 hidden units and 8 heads for each encoder and decoder network. In each encoder and decoder, we have a separate learnable position embedding and its associated word embedding component.

We use a greedy decoding scheme for paraphrase generation, which is fast and cheap to compute. For model optimisation, we use Adam (Kingma and Ba, 2015) as our optimiser with default hyper-parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$). We conduct all the experiments with a batch size of 512 for the Quora and MSCOCO datasets. We set the learning rate as $1e-4$ for MSCOCO and $2e-4$ for Quora based on empirical experiments. All experiments are run for a maximum of 30 epochs on NVidia GPU Cluster with A100 GPU. Experiments are

[6] The paragram-phrase score measures the semantic similarity between a pair of sentences. For the complete PARANMT dataset, the mean score is 0.69, with a standard deviation of 0.26. We select the range among one standard deviation of the mean, i.e. (0.43, 0.95). Since we wish the dataset to be in an actual paraphrase form, we further limit the lower bound to 0.9, to ensure the quality of the data.

[7] The authors do not make their code publicly available.

Table 6
Experiment results for MSCOCO.

Model	B-1	B-2	B-3	B-4	i-B	R-1	R-2	R-L
Upper Bound (Copy Source)	64.97	44.90	30.69	21.30	–	39.18	12.96	34.61
Lower Bound (Random Select)	32.34	10.99	3.81	1.68	–	17.58	1.51	16.27
Residual-LSTM (Prakash et al., 2016)	70.24	48.65	34.04	23.66	18.72	41.07	15.26	37.35
β-VAE (Higgins et al., 2017)	70.04	47.59	32.29	22.54	18.34	40.72	14.75	36.75
Transformer (Vaswani et al., 2017)	71.31	49.86	35.55	24.68	19.81	41.49	15.84	37.09
LBOW-TOPk (Fu et al., 2019)	72.60	51.14	35.66	25.27	21.07	42.08	16.13	38.16
IANet+X (Su et al., 2021)	72.10	52.22	37.39	26.06	21.28	43.81	16.35	39.65
Transformer (our implementation)	68.72	49.64	35.87	26.63	18.59	42.09	16.53	38.35
DDL (our model)	70.75 ↑	51.72 ↑	37.62 ↑	27.95 ↑	19.37 ↑	43.00 ↑	17.01 ↑	39.06 ↑
DDL + SVAR (our model)	70.84 ↑	51.84 ↑	37.75 ↑	28.04 ↑	19.39 ↑	**43.05** ↑	**17.04** ↑	**39.07** ↑
DDL + SVAR* (our model)	**70.99** ↑	**51.91** ↑	**37.82** ↑	**28.12** ↑	**19.39** ↑	43.00 ↑	17.03 ↑	39.02 ↑

repeated three times, with different random seeds (1000, 2000 and 3000), and the average result is reported in Tables 1–6.

6.4. Evaluation

In this paper, we evaluate our models first based on *quantitative metrics*: BLEU (Papineni et al., 2002),[8] ROUGE (Lin, 2004),[9] and i-BLEU (Sun and Zhou, 2012). The main justification behind this is to compare with existing work in the literature directly, which focused on end-to-end learning of paraphrase generation tasks from scratch, using deep neural network models. BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores are based on 'n-gram' coverage between system-generated paraphrase(s) and reference sentences. They have been used widely, to automatically evaluate the quality and accuracy of natural language generation tasks.

Previous work has shown that automatic evaluation metrics can perform well for paraphrase identification tasks (Madnani et al., 2012) and correlate well with human judgements in evaluating generated paraphrases (Wubben et al., 2010). Recent papers introduce additional i-BLEU (Sun and Zhou, 2012) metrics, to balance the fidelity of generated outputs to reference paraphrases (BLEU), as well as the level of diversity introduced (self-B). For all metrics apart from self-B, the higher the value, the better the model performs.

Additionally, we present *qualitative evaluation* results in Tables 9 and 10, based on the Quora dataset. Our qualitative evaluation aims to examine two aspects, as follows: (1) how does our proposed supervised model (DDL) compare with the very strong supervised learning baseline Transformer, given different sizes of labelled pairs data sets; and (2) what are the benefits of our proposed semi-supervised model (DDL+VSAR) when incorporating more unlabelled data, given the labelled pairs data set size remains constant.

6.5. Results and discussion

6.5.1. Learning with unlabelled data only

In an initial experiment, we explored the ability of the VSAR model to perform paraphrase generation tasks using only unlabelled data. This experiment was conducted to see if the model could accurately capture the information required for paraphrase generation, without the aid of labelled data. However, the results of our initial experiment showed that the VSAR model alone was not able to produce high-quality paraphrases, and often resulted in sentences that were either incomprehensible or meaningless. These results motivated us to pursue a semi-supervised learning solution for paraphrase generation, which would provide the model with guidance from labelled data. For unsupervised learning with VSAR, we found that while the lower bound indicated by $\boldsymbol{L}_1$ (Equation (7)) decreased during training in a fully unsupervised setting, the model still generated low-quality paraphrases. This further validated the need for a semi-supervised learning solution, which is introduced in Section 5 of this paper.

Table 7
Complement results for Quora.

Model	B-4	self-B	i-B
Separator (Hosking and Lapata, 2021)	23.68	24.20	14.10
HRQ-VAE (Hosking et al., 2022)	33.11	40.35	18.42
Transformer (our implementation)	26.92	**35.33**	14.47
DDL + SVAR (our model)	28.15 ↑	38.92 ↓	**14.73** ↑
DDL + SVAR* (our model)	**28.16** ↑	39.07 ↓	14.71 ↑

Table 8
Complement results for MSCOCO.

Model	B-4	self-B	i-B
Separator (Hosking and Lapata, 2021)	20.59	12.76	13.92
HRQ-VAE (Hosking et al., 2022)	27.90	16.58	19.04
Transformer (our implementation)	26.87	**13.50**	18.79
DDL + SVAR (our model)	27.87 ↑	15.42 ↓	19.21 ↑
DDL + SVAR* (our model)	**27.92** ↑	15.21 ↓	**19.29** ↑

6.5.2. Learning with a fraction of data

In this section, we present results which are based on a fraction of labelled data in Tables 1, 2 3 and 4. In all four tables, we present the results of two models — the supervised learning model, DDL and the semi-supervised learning model, DDL + VSAR. In a semi-supervised learning setting, VSAR is trained on unlabelled data, and DDL is trained on labelled data. The DDL+VSAR$_1$ model employs equivalent sized labelled and unlabelled datasets, which come from the same source and target pairs, so there is no additional information applied in this case. The DDL+VSAR$_2$ model employs the full unlabelled dataset in addition to the existing labelled dataset, which is the true semi-supervised setting.

Results suggest that the DDL+VSAR$_1$ model achieves competitive or better performance on most metrics' scores compared to the supervised DDL model only trained on labelled data; especially with a lower fraction of the data (for example, the significant improvement for 20K is more noticeable than for 50K). Furthermore, fixing the labelled data size, the DDL+VSAR$_2$ model achieves significantly better performance by using additional unlabelled data, than all other models reported in both tables ($p < .05$; Wilcoxon test), which means the semi-supervised learning does work in this scenario.

6.5.3. Learning with complete data

In this section, we present results based on all labelled data in Tables 5 and 6. Each table comes with three sections. In the first section, we present an upper bound[10] (copying the source as a paraphrase) and

[8] https://www.nltk.org/

[9] https://github.com/huggingface/datasets/tree/master/metrics/rouge

[10] Note, here the upper bound does not refer to performance upper bound, rather than an indicator of satisfaction-level performance.

Table 9
Selected paraphrase generation results for Transformer (TRANS) versus DDL model with different amounts of labelled data (denoted in brackets), represented in the case of Quora dataset.

Source:	what are best courses for journalism ?
Reference:	what are the best courses on journalism ?
TRANS (20K)	which is the best software for beginner ?
DDL (20K)	what is the best digital marketing course ?
TRANS (50K)	what are the best courses on nagpur ?
DDL (50K)	what are the best courses for journalism ?
TRANS (100K)	what are the best courses about journalism ?
DDL (100K)	what are the best courses for journalism ?
Source:	what helps asthma without an inhaler ?
Reference:	what are some ways to help someone with asthma without an inhaler ?
TRANS (20K)	what happens if a range of a range ofr collides with a range ofr 000r
DDL (20K)	how can i save my without doing waves in a month ?
TRANS (50K)	what can be done to work in a ppr ?
DDL (50K)	what are some of the uses an asthma without an inhaler ?
TRANS (100K)	what is the procedure to be an emergency inr ?
DDL (100K)	how can i help asthma without an inhaler ?
Source:	how can i get 1 million users to sign up to my app ?
Reference:	how can i get a million users on my social app ?
TRANS (20K)	how can i get a friend store ?
DDL (20K)	how can i get a game of app on my app ?
TRANS (50K)	how can i get a million million on my startup ?
DDL (50K)	how do i get the first million users app in my app ?
TRANS (100K)	how can i get a million users to write an app ?
DDL (100K)	how do i get a million users to sign up for my app
Source:	can anyone suggest me the best laptop under 35 k in india ?
Reference:	which is the best laptop under 35 , 000 inr ?
TRANS (20K)	what are the best laptop options available for a laptoprsrsrsrsrsrsrsrsrs
DDL (20K)	which is the best laptop to buy in india ?
TRANS (50K)	which is the best laptop under rs . 50000 in india ?
DDL (50K)	what are the best laptops under 35 , 000 in india ?
TRANS (100K)	which is the best laptop to buy under rs . 50000 in india ?
DDL (100K)	which is the best laptop to buy under 35 , 000 in india ?
Source:	how do i manage my microsoft account ?
Reference:	how can i do manage my microsoft account ?
TRANS (20K)	how do i manage my google account ?
DDL (20K)	how do i manage my microsoft office ?
TRANS (50K)	how do i manage my ip address ?
DDL (50K)	how do i manage microsoft microsoft office ?
TRANS (100K)	how do i manage my microsoft account ?
DDL (100K)	how do i manage my microsoft account ?

Table 10
Selected paraphrase generation results for semi-supervised model (DDL+VSAR) when incorporating different amounts of unlabelled data (denoted in brackets) and the same amount of labelled data (20K), represented in the case of Quora dataset.

Source:	is it possible to go to the core of the earth ?
Reference:	if i really wanted to , can i dig all the way to the core of the earth ?
DDL(20K)	is it possible to go to the earth ?
DDL(20K) + VSAR(20K)	how do i go about the earth ?
DDL(20K) + VSAR(50K)	how do i go about the earth ?
DDL(20K) + VSAR(100K)	is it possible to go to the core of the earth ?
Source:	what are best courses for journalism ?
Reference:	what are the best courses on journalism ?
DDL(20K)	what is the best digital marketing course ?
DDL(20K) + VSAR(20K)	what is the best digital marketing agency ?
DDL(20K) + VSAR(50K)	what is the best digital marketing course ?
DDL(20K) + VSAR(100K)	what are the best courses for journalism ?
Source:	how should i stop thinking about someone ?
Reference:	how do i stop thinking about someone ?
DDL(20K)	how do i stop thinking about me ?
DDL(20K) + VSAR(20K)	how do i stop thinking about thinking ?
DDL(20K) + VSAR(50K)	how do i stop thinking about something ?
DDL(20K) + VSAR(100K)	how do i stop thinking about someone ?
Source:	what motivates all people ?
Reference:	what motivates people ?
DDL(20K)	why do people often keep all people ?
DDL(20K) + VSAR(20K)	why do people get tattoos ?
DDL(20K) + VSAR(50K)	what inspires to be so hard ?
DDL(20K) + VSAR(100K)	what motivates people ?
Source:	how can i get 1 million users to sign up to my app ?
Reference:	how can i get a million users on my social app ?
DDL(20K)	how can i get a game of app on my app ?
DDL(20K) + VSAR(20K)	how do i get a person from a app ?
DDL(20K) + VSAR(50K)	how can i get a billionaire by youtube ?
DDL(20K) + VSAR(100K)	how can i get 1 million users back from my app ?
Source:	is vegetarian good for health or non - vegetarian ?
Reference:	which is good food for our health : vegetarian or non - vegetarian ?
DDL(20K)	is smoking considered a vegetarian vegetarian ?
DDL(20K) + VSAR(20K)	is vegetarian considered good for health ?
DDL(20K) + VSAR(50K)	is vegetarian better than vegetarian ?
DDL(20K) + VSAR(100K)	is vegetarian health good or bad ?
Source:	how do i manage my microsoft account ?
Reference:	how can i do manage my microsoft account ?
DDL(20K)	how do i manage my microsoft office ?
DDL(20K) + VSAR(20K)	how do i manage my microsoft size ?
DDL(20K) + VSAR(50K)	how do i manage my google account ?
DDL(20K) + VSAR(100K)	how do i manage my microsoft account ?

a lower bound (randomly selecting the ground truth as a paraphrase) calculated based on the test split (as in Chen et al. (2020)). This is used as an indication of how well the model performs. In the second section, we present major state-of-the-art models published in recent years. In the third section, we present our own implementation of the Transformer model, which we consider as our absolute baseline, and present results for our models. Our implementation is competitive with the ones reported in recent papers. For our models, DDL is our supervised model, DDL+VSAR is our semi-supervised model, and DDL+VSAR* is our model with no prior used. Compared with state-of-the-art supervised models, our models, in general, achieve statistically significant better BLEU scores and competitive Rouge scores for both datasets ($p < .05$; Wilcoxon test). Our complementary experimental results are presented in Tables 7 and 8, which we compare with two more recent state-of-the-art models. Our models once again achieve statistically significant better or competitive performance than the reported ($p < .05$; Wilcoxon test), which means our semi-supervised model is competitive with state-of-the-art supervised baselines.

6.5.4. Qualitative evaluation for supervised learning with labelled data

In Table 9, we present examples from the Quora test data set and their corresponding model outputs from our proposed supervised learning model DDL (introduced in Section 4) and outputs from a very strong baseline model Transformer (denoted as TRANS), using varying amounts of training data. The table first presents the source and golden reference pair, followed by the outputs of the models (DDL and TRANS) trained on 20K, 50K, and 100K labelled dataset pairs. Each example was generated based on a random seed setting of 1000, ensuring a fair qualitative evaluation; additionally, we always make sure that a smaller amount of labelled pairs is a subset of examples from a larger data size group, this allows us to better quantify the benefits of adding more unlabelled data.

It is quite clear from the results that the generated paraphrase is more accurate in terms of semantic information, and that it matches better with the reference, when more labelled data is used. Although, at the same time, we observe that the advantages of DDL become more significant when the number of labelled pairs is scarce (i.e. with 20K, the improvement is more significant than with 50K and 100K). Additionally, our DDL model demonstrates a clear advantage over the TRANS model in capturing the essence of the information, as seen in its ability to capture critical details (e.g. capture asthma and inhaler in the second example; capture number 35000 instead of 50000 in the fourth example). The DDL model also showed more efficient learning, as it was able to achieve comparable results using 50K data, as opposed to 100K data for the TRANS model in several examples. These observations further reinforce the effectiveness and efficiency of our proposed DDL model.

6.5.5. Qualitative evaluation for incorporating unlabelled data

In Table 10, we present qualitative examples from the Quora test data set and their corresponding model outputs from our proposed semi-supervised learning model DDL + VSAR (from Section 5) given the same amount of labelled data (20K, same data instance as in Section 6.5.4) plus difference size of unlabelled data (20K, 50K and 100K). For comparison, model output with DDL (Section 4), trained with the same 20K examples, is provided. Similarly, as in Table 9, we use the same random seed of 1000 to generate these examples. In Table 10, we first presented the source and golden reference, followed by model outputs trained based on 20K labelled pairs by the DDL model, the same 20K pairs used for the DDL+VSAR model (similar to DDL+VSAR$_1$ setting in Tables 1, 2, 3 and 4), the same 20K with 50K unlabelled data (extra 30K) for the DDL+VSAR model and the same 20K with 100K unlabelled data (extra 80K) for the DDL+VSAR model (similar to DDL+VSAR$_2$ setting in Tables 1, 2, 3 and 4).

We can clearly observe that the generated paraphrase matches better with the reference when more unlabelled data is utilised during the training of the DDL+VSAR model. The latent sequence generated when incorporating unlabelled data in the VSAR model improve the paraphrase generation performance. In general, we observe a progressive improvement in capturing the essence of information when incorporating more unlabelled examples. Compared to the preliminary experiment where the VSAR model failed to learn in a fully unsupervised scenario (Section 6.5.1); through qualitative results in Table 10 and quantitative results in Tables 1, 2, 3 and 4; we show that we are able to conduct semi-supervised learning by combining our unsupervised model VSAR and our supervised model DDL.

6.5.6. Error analysis

In this section, we present an error analysis of our proposed model. In general, our model works well for semi-supervised settings, as demonstrated in Tables 1,2, 3 and 4, however, there are cases that the generated examples from our models get worth quality or no improvements in terms of quality when more unlabelled pairs are utilised. We find that in cases where the DDL+VSAR model has no improvement when seeing more unlabelled data, the golden reference of test examples is relatively easy to paraphrase and is presented in shorter sentence form. However, we also observed scenarios when our DDL+VSAR model overcomplex paraphrases when more unlabelled data points are used, as shown in Table 11. We notice that observing unlabelled data increase the model's ability to handle more complex paraphrase, however, at the same time encourage the model to return a complex answer. This is an interesting discovery and suggests that the quality of unlabelled data also plays an important role in the semi-supervised learning process.

6.5.7. Algorithm run time

In this section, we further discuss the algorithm run time and GPU memory requirement of our proposed models. Regarding the DDL model, it is equivalent to training a standard transformer model with additional cost terms (from two directions: source to target and target to source). Hence, training a DDL model does not require additional run time and requires no additional GPU memory. Regarding the VSAR model alone, since we are sharing the parameters for source reconstruction and latent inference (as shown in Fig. 1), the cost of training is similar to the DDL model, plus some additional GPU memory cost due to the need of saving gradient from latent samples in order to perform back-propagation. Since a single sample is used during our training, the extra GPU memory cost is not significant, and we do not recognise the extra run time for training the VSAR model. Regarding the semi-supervised learning model, DDL+VSAR, we need to first pre-train a DDL model from the labelled data, then initialise the model weight from the DDL model and jointly fine-tune the DDL+VSAR model from both labelled and unlabelled data. The total run time is doubled compared to the process of pre-training and fine-tuning. Although we find that pre-training already set the model at a near-optimal parameter space, and hence we could potentially use fewer epochs for the semi-supervised learning to ensure the algorithm scale well to large-scale data.

7. Limitations and future work

Here we briefly discuss two main limitations which are identified in this research with each associated future work. The first limitation involves the quantitative evaluation metrics, such as BLEU and Rouge score, used in this research. These metrics are based on the overlap of n-gram contexts between generated outputs and reference text, and while they are commonly used to compare with published results (as shown in Tables 5, 6, 7 and 8) , they cannot directly assess the quality of the generated text. More recent evaluation metrics, such as BERTscore (Zhang et al., 2020), have been proposed, but there is still no universal agreement on what are the best quantitative measurements. One future work direction is to look into this area and propose better evaluation metrics for capturing the quality of generated paraphrases, which will ultimately create a meaningful impact on natural language generation research.

The second limitation is the absence of user-based studies to evaluate systems implemented based on our proposed methods, which are often considered a more comprehensive measure of the model performance. Although in this paper, we observed promising results in Section 6.5. Current evaluations in this work rely on public benchmark datasets, which may be biased towards the subset of data the model is trained on. Due to resource constraints, user-based evaluations were

Table 11
Selected paraphrase generation results for error analysis, represented in the case of Quora dataset.

Source:	what is web marketing ? i want to learn internet marketing ?
Reference:	what ' s the best way to get started in internet marketing ?
DDL(20K) + VSAR(20K)	what is the most valuable website for the internet ?
DDL(20K) + VSAR(50K)	what is the best way to learn internet ?
DDL(20K) + VSAR(100K)	what is the best digital marketing course available online and offline in india and why ?
Source:	what do you feel is the purpose of life ?
Reference:	from your perspective , what is the purpose of life ?
DDL(20K) + VSAR(20K)	what is the purpose of life ?
DDL(20K) + VSAR(50K)	what is the purpose of your life ?
DDL(20K) + VSAR(100K)	what do you think is the purpose of life ?
Source:	why do you believe in ghosts and the supernatural ?
Reference:	do you believe in ghosts , why ?
DDL(20K) + VSAR(20K)	do you believe in ghost ?
DDL(20K) + VSAR(50K)	do you believe in ghosts ?
DDL(20K) + VSAR(100K)	do you think you ' ve ever found in a supernatural event ?
Source:	why do so many people post questions on quora that could be easily and thoroughly answered by simply typing the question into any search engine ?
Reference:	why do so many people ask soppy questions on quora , which can easily be googled ?
DDL(20K) + VSAR(20K)	why do people ask questions on quora instead of googling it ?
DDL(20K) + VSAR(50K)	why do people ask questions on quora instead of googling it ?
DDL(20K) + VSAR(100K)	why do people ask questions on quora that can be answered with a quick web search ?
Source:	where can i find delicious cupcakes at gold coast ?
Reference:	where can i found different cupcake flavors in gold coast ?
DDL(20K) + VSAR(20K)	where can i get best quality gourmet cupcakes in gold coast ?
DDL(20K) + VSAR(50K)	where can i get best quality cupcakes in gold coast ?
DDL(20K) + VSAR(100K)	where can i get best flavors , designs and decorations for cupcakes at gold coast ?
Source:	is vegetarian good for health or non - vegetarian ?
Reference:	which is good food for our health : vegetarian or non - vegetarian ?
DDL(20K)	is smoking considered a vegetarian vegetarian ?
DDL(20K) + VSAR(20K)	is vegetarian considered good for health ?
DDL(20K) + VSAR(50K)	is vegetarian better than vegetarian ?
DDL(20K) + VSAR(100K)	is vegetarian health good or bad ?

not performed in this research and represented a promising area for future study.

For future work, in this paper, we have explored a more structured approach to modelling unobserved paraphrase as a discrete latent variable. A future direction worth exploring is to apply and extend this technique in pre-trained language models (PLMs). Recent advancements in PLMs suggest that they can achieve very good performance given a relatively small proportion of data or even performs well in few-shot or zero-shot scenarios in document retrieval tasks. However, challenges still exist in the cases of natural language generation (NLG) tasks, and our proposed method in this paper can serve as a stepping stone to using both labelled and unlabelled data for NLG tasks.

8. Conclusions

In this paper, we have introduced a semi-supervised deep generative model for paraphrase generation. The unsupervised model (VSAR) is based on the variational auto-encoding framework and provides an effective method to handle missing labels. The supervised model (DDL) conducts dual learning and injects supervised information into the unsupervised model. With our novel knowledge-reinforced-learning (KRL) scheme, we empirically demonstrate that semi-supervised learning benefits our combined model, given unlabelled data and a fraction of the paired data. The evaluation results show that our combined model improves upon a very strong baseline model in a semi-supervised setting. We also observe that, even for the full dataset, our combined model achieves competitive performance with the state-of-the-art models for two paraphrase generation benchmark datasets. Additionally, we are able to model language as a discrete latent variable sequence for paraphrase generation tasks. Importantly, the resultant generative model is able to exploit both supervised and unsupervised data in sequence-to-sequence tasks effectively.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. This research was supported by the Durham Doctoral Studentship (DDS) from Durham University awared to Jialin Yu and funded internally by Prof. Alexandra Cristea.

References

Balin, M.F., Abid, A., Zou, J.Y., 2019. Concrete autoencoders: Differentiable feature selection and reconstruction. In: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning. ICML 2019, 9–15 June 2019, Long Beach, California, USA, In: Proceedings of Machine Learning Research, vol. 97, PMLR, pp. 444–453, URL: http://proceedings.mlr.press/v97/balin19a.html.

Bengio, Y., Léonard, N., Courville, A., 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432.

Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., Bengio, S., 2016. Generating sentences from a continuous space. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Association for Computational Linguistics, Berlin, Germany, pp. 10–21. http://dx.doi.org/10.18653/v1/K16-1002, URL: https://aclanthology.org/K16-1002.

Burrows, S., Potthast, M., Stein, B., 2013. Paraphrase acquisition via crowdsourcing and machine learning. ACM Trans. Intell. Syst. Technol. 4 (3), 1–21.

Cao, Z., Luo, C., Li, W., Li, S., 2017. Joint copying and restricted generation for paraphrase. In: Singh, S.P., Markovitch, S. (Eds.), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. February 4–9, 2017, San Francisco, California, USA, AAAI Press, pp. 3152–3158, URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14527.

Chang, E., Demberg, V., Marin, A., 2021. Jointly improving language understanding and generation with quality-weighted weak supervision of automatic labeling. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, pp. 818–829, Online, URL: https://aclanthology.org/2021.eacl-main.69.

Chen, Y., Jiang, H., Liu, L., Wang, R., Shi, S., Xu, R., 2022. MCPG: A flexible multi-level controllable framework for unsupervised paraphrase generation. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 5948–5958.

Chen, J., Song, L., Wainwright, M.J., Jordan, M.I., 2018. Learning to explain: An information-theoretic perspective on model interpretation. In: Dy, J.G., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, In: Proceedings of Machine Learning Research, vol. 80, PMLR, pp. 882–891, URL: http://proceedings.mlr.press/v80/chen18j.html.

Chen, W., Tian, J., Xiao, L., He, H., Jin, Y., 2020. A semantically consistent and syntactically variational encoder-decoder framework for paraphrase generation. In: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 1186–1198. http://dx.doi.org/10.18653/v1/2020.coling-main.102, URL: https://aclanthology.org/2020.coling-main.102.

Choi, J., Yoo, K.M., goo Lee, S., 2018. Learning to compose task-specific tree structures. In: McIlraith, S.A., Weinberger, K.Q. (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. EAAI-18, New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, pp. 5094–5101, URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16682.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423, URL: https://aclanthology.org/N19-1423.

Dieng, A.B., Kim, Y., Rush, A.M., Blei, D.M., 2019. Avoiding latent variable collapse with generative skip models. In: Chaudhuri, K., Sugiyama, M. (Eds.), The 22nd International Conference on Artificial Intelligence and Statistics. AISTATS 2019, 16–18 April 2019, Naha, Okinawa, Japan, In: Proceedings of Machine Learning Research, 89, PMLR, pp. 2397–2405, URL: http://proceedings.mlr.press/v89/dieng19a.html.

Dong, L., Mallinson, J., Reddy, S., Lapata, M., 2017. Learning to paraphrase for question answering. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark, pp. 875–886. http://dx.doi.org/10.18653/v1/D17-1091, URL: https://aclanthology.org/D17-1091.

Du, J., Grave, E., Gunel, B., Chaudhary, V., Celebi, O., Auli, M., Stoyanov, V., Conneau, A., 2021. Self-training improves pre-training for natural language understanding. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pp. 5408–5418. http://dx.doi.org/10.18653/v1/2021.naacl-main.426, Online, URL: https://aclanthology.org/2021.naacl-main.426.

Fu, Y., Feng, Y., Cunningham, J.P., 2019. Paraphrase generation with latent bag of words. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019. NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, pp. 13623–13634, URL: https://proceedings.neurips.cc/paper/2019/hash/5e2b66750529d8ae895ad2591118466f-Abstract.html.

Guo, H., Pasunuru, R., Bansal, M., 2018a. Dynamic multi-level multi-task learning for sentence simplification. In: Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 462–476, URL: https://aclanthology.org/C18-1039.

Guo, H., Pasunuru, R., Bansal, M., 2018b. Soft layer-specific multi-task summarization with entailment and question generation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp. 687–697. http://dx.doi.org/10.18653/v1/P18-1064, URL: https://aclanthology.org/P18-1064.

Gupta, A., Agarwal, A., Singh, P., Rai, P., 2018. A deep generative framework for paraphrase generation. In: McIlraith, S.A., Weinberger, K.Q. (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. EAAI-18, New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, pp. 5149–5156, URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16353.

He, J., Spokoyny, D., Neubig, G., Berg-Kirkpatrick, T., 2019. Lagging inference networks and posterior collapse in variational autoencoders. In: 7th International Conference on Learning Representations. ICLR 2019, New Orleans, la, USA, May 6–9, 2019, OpenReview.net, URL: https://openreview.net/forum?id=rylDfnCqF7.

He, J., Wang, X., Neubig, G., Berg-Kirkpatrick, T., 2020. A probabilistic formulation of unsupervised text style transfer. In: 8th International Conference on Learning Representations. ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, URL: https://openreview.net/forum?id=HJlA0C4tPS.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., Ma, W., 2016. Dual learning for machine translation. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016. December 5–10, 2016, Barcelona, Spain, pp. 820–828, URL: https://proceedings.neurips.cc/paper/2016/hash/5b69b9cb83065d403869739ae7f0995e-Abstract.html.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A., 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. OpenReview.net, URL: https://openreview.net/forum?id=Sy2fzU9gl.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507.

Hosking, T., Lapata, M., 2021. Factorising meaning and form for intent-preserving paraphrasing. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, pp. 1405–1418. http://dx.doi.org/10.18653/v1/2021.acl-long.112, Online, URL: https://aclanthology.org/2021.acl-long.112.

Hosking, T., Tang, H., Lapata, M., 2022. Hierarchical sketch induction for paraphrase generation. arXiv preprint arXiv:2203.03463.

Hu, J.E., Rudinger, R., Post, M., Van Durme, B., 2019. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01. pp. 6521–6528.

Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., Zhao, Z., 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. arXiv preprint arXiv:2301.12661.

Jang, E., Gu, S., Poole, B., 2017. Categorical reparameterization with gumbel-softmax. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. OpenReview.net, URL: https://openreview.net/forum?id=rkE3y85ee.

Kauchak, D., Barzilay, R., 2006. Paraphrasing for automatic evaluation. In: Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. Association for Computational Linguistics, New York City, USA, pp. 455–462, URL: https://aclanthology.org/N06-1058.

Kim, Y., Wiseman, S., Rush, A.M., 2018. A tutorial on deep latent variable models of natural language. arXiv preprint arXiv:1812.06834.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings. URL: http://arxiv.org/abs/1412.6980.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational Bayes. In: Bengio, Y., LeCun, Y. (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings. URL: http://arxiv.org/abs/1312.6114.

Kumar, A., Ahuja, K., Vadapalli, R., Talukdar, P., 2020. Syntax-guided controlled generation of paraphrases. Trans. Assoc. Comput. Linguist. 8, 329–345. http://dx.doi.org/10.1162/tacl_a_00318, URL: https://aclanthology.org/2020.tacl-1.22.

Lee, M., Cimino, J., Zhu, H.R., Sable, C., Shanker, V., Ely, J., Yu, H., 2006. Beyond information retrieval—medical question answering. In: AMIA Annual Symposium Proceedings, Vol. 2006. American Medical Informatics Association, p. 469.

Li, Z., Jiang, X., Shang, L., Liu, Q., 2019. Decomposable neural paraphrase generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 3403–3414. http://dx.doi.org/10.18653/v1/P19-1332, URL: https://aclanthology.org/P19-1332.

Lin, C.-Y., 2004. ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain, pp. 74–81, URL: https://aclanthology.org/W04-1013.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

Liu, Y., Li, S., Cao, Y., Lin, C.-Y., Han, D., Yu, Y., 2008. Understanding and summarizing answers in community-based question answering services. In: Proceedings of the 22nd International Conference on Computational Linguistics. Coling 2008, Coling 2008 Organizing Committee, Manchester, UK, pp. 497–504, URL: https://aclanthology.org/C08-1063.

Luong, M., Le, Q.V., Sutskever, I., Vinyals, O., Kaiser, L., 2016. Multi-task sequence to sequence learning. In: Bengio, Y., LeCun, Y. (Eds.), 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings. URL: http://arxiv.org/abs/1511.06114.

Maddison, C.J., Mnih, A., Teh, Y.W., 2017. The concrete distribution: A continuous relaxation of discrete random variables. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. OpenReview.net, URL: https://openreview.net/forum?id=S1jE5L5gl.

Madnani, N., Tetreault, J., Chodorow, M., 2012. Re-examining machine translation metrics for paraphrase identification. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Montréal, Canada, pp. 182–190, URL: https://aclanthology.org/N12-1019.

Martins, A.F.T., Mihaylova, T., Nangia, N., Niculae, V., 2019. Latent structure models for natural language processing. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts. Association for Computational Linguistics, Florence, Italy, pp. 1–5. http://dx.doi.org/10.18653/v1/P19-4001, URL: https://aclanthology.org/P19-4001.

Meng, Y., Ao, X., He, Q., Sun, X., Han, Q., Wu, F., Li, J., et al., 2021. ConRPG: Paraphrase generation using contexts as regularizer. arXiv preprint arXiv:2109.00363.

Mercatali, G., Freitas, A., 2021. Disentangling generative factors in natural language with discrete variational autoencoders. arXiv preprint arXiv:2109.07169.

Miao, Y., Blunsom, P., 2016. Language as a latent variable: Discrete generative models for sentence compression. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pp. 319–328. http://dx.doi.org/10.18653/v1/D16-1031, URL: https://aclanthology.org/D16-1031.

Miao, Y., Yu, L., Blunsom, P., 2016. Neural variational inference for text processing. In: Balcan, M.-F., Weinberger, K.Q. (Eds.), Proceedings of the 33nd International Conference on Machine Learning. ICML 2016, New York City, NY, USA, June 19–24, 2016, In: JMLR Workshop and Conference Proceedings, vol. 48, JMLR.org, pp. 1727–1736, URL: http://proceedings.mlr.press/v48/miao16.html.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a Meeting. December 5–8, 2013, Lake Tahoe, Nevada, United States, pp. 3111–3119, URL: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Miller, G.A., 1992. WordNet: A lexical database for english. In: Speech and Natural Language: Proceedings of a Workshop. Harriman, New York, February 23–26, 1992, URL: https://aclanthology.org/H92-1116.

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267, 1–38.

Mnih, A., Gregor, K., 2014. Neural variational inference and learning in belief networks. In: Proceedings of the 31th International Conference on Machine Learning. ICML 2014, Beijing, China, 21–26 June 2014, In: JMLR Workshop and Conference Proceedings, vol. 32, JMLR.org, pp. 1791–1799, URL: http://proceedings.mlr.press/v32/mnih14.html.

Mnih, V., Heess, N., Graves, A., Kavukcuoglu, K., 2014. Recurrent models of visual attention. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014. December 8–13 2014, Montreal, Quebec, Canada, pp. 2204–2212, URL: https://proceedings.neurips.cc/paper/2014/hash/09c6c3783b4a70054da74f2538ed47c6-Abstract.html.

Narayan, S., Reddy, S., Cohen, S.B., 2016. Paraphrase generation from latent-variable PCFGs for semantic parsing. In: Proceedings of the 9th International Natural Language Generation Conference. Association for Computational Linguistics, Edinburgh, UK, pp. 153–162. http://dx.doi.org/10.18653/v1/W16-6625, URL: https://aclanthology.org/W16-6625.

Niculae, V., Corro, C.F., Nangia, N., Mihaylova, T., Martins, A.F., 2023. Discrete latent structure in neural networks. arXiv preprint arXiv:2301.07473.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. http://dx.doi.org/10.3115/1073083.1073135, URL: https://aclanthology.org/P02-1040.

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., Callison-Burch, C., 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, Beijing, China, pp. 425–430. http://dx.doi.org/10.3115/v1/P15-2070, URL: https://aclanthology.org/P15-2070.

Pennington, J., Socher, R., Manning, C., 2014. GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP, Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543. http://dx.doi.org/10.3115/v1/D14-1162, URL: https://aclanthology.org/D14-1162.

Prakash, A., Hasan, S.A., Lee, K., Datla, V., Qadir, A., Liu, J., Farri, O., 2016. Neural paraphrase generation with stacked residual LSTM networks. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. The COLING 2016 Organizing Committee, Osaka, Japan, pp. 2923–2934, URL: https://aclanthology.org/C16-1275.

Razavi, A., van den Oord, A., Poole, B., Vinyals, O., 2019. Preventing posterior collapse with delta-VAEs. In: 7th International Conference on Learning Representations. ICLR 2019, New Orleans, la, USA, May 6–9, 2019, OpenReview.net, URL: https://openreview.net/forum?id=BJe0Gn0cY7.

Rezende, D.J., Mohamed, S., Wierstra, D., 2014. Stochastic backpropagation and approximate inference in deep generative models. In: Proceedings of the 31th International Conference on Machine Learning. ICML 2014, Beijing, China, 21–26 June 2014, In: JMLR Workshop and Conference Proceedings, vol. 32, JMLR.org, pp. 1278–1286, URL: http://proceedings.mlr.press/v32/rezende14.html.

Schneider, F., Jin, Z., Schölkopf, B., 2023. Moûsai: Text-to-music generation with long-context latent diffusion. arXiv preprint arXiv:2301.11757.

Su, S.-Y., Chuang, Y.-S., Chen, Y.-N., 2020a. Dual inference for improving language understanding and generation. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, pp. 4930–4936. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.443, Online, URL: https://aclanthology.org/2020.findings-emnlp.443.

Su, S.-Y., Huang, C.-W., Chen, Y.-N., 2019. Dual supervised learning for natural language understanding and generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 5472–5477. http://dx.doi.org/10.18653/v1/P19-1545, URL: https://aclanthology.org/P19-1545.

Su, S.-Y., Huang, C.-W., Chen, Y.-N., 2020b. Towards unsupervised language understanding and generation by joint dual learning. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 671–680. http://dx.doi.org/10.18653/v1/2020.acl-main.63, Online, URL: https://aclanthology.org/2020.acl-main.63.

Su, Y., Vandyke, D., Baker, S., Wang, Y., Collier, N., 2021. Keep the primary, rewrite the secondary: A two-stage approach for paraphrase generation. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, pp. 560–569. http://dx.doi.org/10.18653/v1/2021.findings-acl.50, Online, URL: https://aclanthology.org/2021.findings-acl.50.

Sun, H., Zhou, M., 2012. Joint learning of a dual SMT system for paraphrase generation. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Jeju Island, Korea, pp. 38–42, URL: https://aclanthology.org/P12-2008.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014. December 8–13 2014, Montreal, Quebec, Canada, pp. 3104–3112, URL: https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

Van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Mach. Learn. 109 (2), 373–440.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. December 4–9, 2017, Long Beach, CA, USA, pp. 5998–6008, URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, Y., Blei, D., Cunningham, J.P., 2021. Posterior collapse and latent variable non-identifiability. Adv. Neural Inf. Process. Syst. 34.

Wang, Y., Zhai, C., Hassan, H., 2020. Multi-task learning for multilingual neural machine translation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. EMNLP, Association for Computational Linguistics, pp. 1022–1034. http://dx.doi.org/10.18653/v1/2020.emnlp-main.75, Online, URL: https://aclanthology.org/2020.emnlp-main.75.

Wen, T., Miao, Y., Blunsom, P., Young, S.J., 2017. Latent intention dialogue models. In: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning. ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, In: Proceedings of Machine Learning Research, vol. 70, PMLR, pp. 3732–3741, URL: http://proceedings.mlr.press/v70/wen17a.html.

Wieting, J., Gimpel, K., 2017. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. arXiv preprint arXiv:1711.05732.

Williams, R.J., Zipser, D., 1989. A learning algorithm for continually running fully recurrent neural networks. Neural Comput. 1 (2), 270–280.

Wubben, S., van den Bosch, A., Krahmer, E., 2010. Paraphrase generation as monolingual translation: Data and evaluation. In: Proceedings of the 6th International Natural Language Generation Conference. URL: https://aclanthology.org/W10-4223.

Xie, J., Chen, J., Cai, Y., Huang, Q., Li, Q., 2023. Visual paraphrase generation with key information retained. ACM Trans. Multimed. Comput. Commun. Appl.

Xie, S.M., Ermon, S., 2019. Reparameterizable subset sampling via continuous relaxations. In: Kraus, S. (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. IJCAI 2019, Macao, China, August 10–16, 2019, ijcai.org, pp. 3919–3925. http://dx.doi.org/10.24963/ijcai.2019/544, URL: https://doi.org/10.24963/ijcai.2019/544.

Yang, Z., Hu, Z., Dyer, C., Xing, E.P., Berg-Kirkpatrick, T., 2018. Unsupervised text style transfer using language models as discriminators. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018. NeurIPS 2018, December 3–8, 2018, Montréal, Canada, pp. 7298–7309, URL: https://proceedings.neurips.cc/paper/2018/hash/398475c83b47075e8897a083e97eb9f0-Abstract.html.

Yao, X., Van Durme, B., 2014. Information extraction over structured data: Question answering with freebase. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Baltimore, Maryland, pp. 956–966. http://dx.doi.org/10.3115/v1/P14-1090, URL: https://aclanthology.org/P14-1090.

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2020. BERTScore: Evaluating Text Generation with BERT, International Conference on Learning Representations.

Zhou, J., Bhat, S., 2021. Paraphrase generation: A survey of the state of the art. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 5075–5086.

Zhu, X.J., 2005. Semi-Supervised Learning Literature Survey. University of Wisconsin-Madison Department of Computer Sciences.