Improving Machine Translation and Summarization with the Sinkhorn Divergence

Shijie Li[1], Inigo Jauregi Unanue[1,2], and Massimo Piccardi[1]

[1] University of Technology Sydney, New South Wales, Australia
Shijie.Li@student.uts.edu.au, Massimo.Piccardi@uts.edu.au
[2] RoZetta Technology, Sydney, New South Wales, Australia
Inigo.Jauregi@rozettatechnology.com

Abstract. Important natural language processing tasks such as machine translation and document summarization have made enormous strides in recent years. However, their performance is still partially limited by the standard training objectives, which operate on single tokens rather than on more global features. Moreover, such standard objectives do not explicitly consider the source documents, potentially affecting their alignment with the predictions. For these reasons, in this paper, we propose using an Optimal Transport (OT) training objective to promote a global alignment between the model's predictions and the source documents. In addition, we present an original implementation of the OT objective based on the Sinkhorn divergence between the final hidden states of the model's encoder and decoder. Experimental results over machine translation and abstractive summarization tasks show that the proposed approach has been able to achieve statistically significant improvements across all experimental settings compared to our baseline and other alternative objectives. A qualitative analysis of the results also shows that the predictions have been able to better align with the source sentences thanks to the supervision of the proposed objective.

Keywords: Natural Language Processing · Natural Language Generation · Neural Text Generation · Optimal Transport

1 Introduction

Natural language generation (NLG), a key field for the natural language processing (NLP) community, lends itself to a wide range of applications such as machine translation, text summarization, dialogue systems, and others [14]. In these tasks, attention-based sequence-to-sequence (seq2seq) models [23] are dominant, together with the conventional maximum-likelihood estimation (MLE), which is also known as *teacher forcing* in the area of recurrent neural networks (RNN). This approach maximizes the generation probability of the current target word conditioned on all the previous ground-truth inputs and all the source words. However, it has been widely criticized for both its conditioning of the predictions on ground-truth information, unavailable at inference time, and its inability to capture sentence-level features by only operating at token level [19].

Several attempts have been made to address the limitations of standard MLE by adopting sentence-level objectives. For example, Ranzato et al. [19] have trained their models by directly optimizing sentence-level evaluation metrics such as the BLEU [16] and ROUGE [13] scores. However, these two metrics compute sentence similarity primarily based on surface matches of n-grams. While they can provide sentence-level information to a certain extent, they struggle to reward context-preserving lexical equivalence. Additionally, they are typically based on *hard* predictions, i.e., sequences of labels, which make the metrics non-differentiable as they are flat and subject to change discontinuously in the parameter space. For this reason, optimizing them often requires resorting to slow and high-variance gradient estimation techniques such as policy gradient.

An efficient alternative to optimizing the above n-gram-based metrics is to optimize embedding-based ones. Several such metrics have been proposed in recent years, including the Word Mover's Distance [10], MoverScore [26], and BERTScore [25]. The matching schemes in these advanced sentence-level metrics better preserve semantically-relevant information, especially when combined with contemporary pretrained language models such as BERT [5] and BART [11]. More importantly, they straightforwardly support optimization as they are based on continuous quantities. For instance, Jauregi Unanue et al. [8] have proposed fine-tuning machine translation models by using BERTScore as the training objective, reporting consistent improvements over a variety of language pairs. However, most of these methods only focus on measuring the similarity between predictions and references, and rarely pay attention to the source documents. While the source information can be covered by the inner cross-attention mechanism to some extent, the attention mechanism itself has been criticized in recent years. For example, in the absence of constraints, some of the source tokens may be rarely attended to. To amend this, some approaches have started to include explicit coverage terms in the models [7,17].

To address the above issues, in this paper, we focus on providing text generation models with sentence-level supervision directly from the source text. To achieve this goal, we propose a novel training objective based on the minimization of the recently-proposed Sinkhorn divergence (SD) [6] between the hidden states of the encoder and decoder. The Sinkhorn divergence is a variant of the general optimal transport (OT) problem, which can be used to optimally align two arbitrary sets of weighted elements. The proposed objective only utilizes the contextualized source information already learned by the encoder, without introducing any additional module or memory footprint. In addition, the inference remains unchanged and its run time is unaffected. Overall, our paper makes the following main contributions:

- A novel training objective for conditional text generation models such as machine translation and document summarization providing sentence-level supervision directly from the source text.
- An original implementation of the objective leveraging the context-aware hidden states of the encoder and the decoder, and the Sinkhorn divergence – a performing variant of OT distance.
- Experimental results on machine translation and abstractive summarization showing marked improvements in both text quality and word alignment over an MLE baseline and all other compared objectives.

2 Related Work

Sentence-Level Supervision The sentence-level supervision used in early research was typically performed with the non-differentiable metrics used for evaluation, such as the BLEU and ROUGE scores used in [19]. With the increases in model capacity and training data size, advanced language models such as BERT [5] and BART [11] have shown their ability to learn context-aware representations of the input sentences. As a result, researchers have started to leverage these pretrained representations as sentence-and document-level signals. For example, Zhang et al. [25] have utilized BERT as a sentence-level evaluation metric, and Chen et al. [3] have focused on distilling knowledge learned by a large BERT model for training smaller, student models. Typically, these context-aware representations are extracted from the last layer of the language models, and we follow this line in our implementation.

Coverage of Source Information Approaches for explicitly covering the source-side information are an important component of statistical machine translation, and also the founding idea behind the attention mechanism in contemporary NLG models [24], which adaptively focus on different parts of the source sentence at each generation step. In addition, networks such as the copying net [28] have directly allowed copying content from the source text to the predictions, leveraging the homogeneity of the source information and the output in NLG tasks. More recently, Garg et al. [7] have jointly trained an explicit alignment module for source and target sentences when training machine translation models and Parnell et al. [17] have proposed a reinforcement learning reward for multi-document summarization to even out the individual contributions of the source documents.

Optimal Transport Optimal transport (OT) was first introduced in NLP by Kusner et al. [10] as a way to measure the distance between two documents. Since then, OT has been widely used in several other applications. For instance, Alqahtani et al. [1] have utilized it as an objective for word alignment while Zhao et al. [26] have used it as an evaluation metric. In terms of NLG tasks, Chen et al. [2] and Wang et al. [12] have demonstrated improved performance by minimizing the OT distance between the references and sentences generated with teacher forcing (TFOT) and student forcing (SFOT), respectively. In turn, Nguyen et al.[15] have used the OT distance for knowledge distillation and shown improvements in cross-lingual text summarization. However, none of these OT-based methods has paid explicit attention to context-aware source information.

3 Methodology

In this section, we first briefly recap the standard seq2seq training, then provide the basics of OT optimization, and finally introduce the proposed approach.

3.1 Sequence-to-Sequence Model Training

The seq2seq framework is essentially an encoder-decoder architecture, where the encoder is responsible for mapping a source sentence $X_{1:N} = (x_1, \dots, x_N)$ to a sequence of hidden vectors, or states, $H_{1:N} = (h_1, \ldots, h_N)$, and the decoder is responsible for eventually mapping these hidden vectors to a target sentence, $Y_{1:M} = (y_1, \ldots, y_M)$. In the original seq2seq model [23], only the last hidden state of the encoder, h_N, was fed into the decoder, limiting the source information available to the decoder. However, this limitation was removed by the attention mechanisms [14], which leverage all the encoder's hidden states. The standard MLE training objective of seq2seq models is to minimize the negative log-likelihood of the target sentence, $Y_{1:M}$, conditioned on $X_{1:N}$:

$$\mathcal{L}_{\mathrm{MLE}} = -\log P_\theta(Y_{1:M}|X_{1:N}) = -\sum_{m=1}^{M} \log P_\theta(y_m|y_{<m}, X_{1:N}) \tag{1}$$

3.2 The Proposed Approach: a Contextual Sinkhorn Divergence

Optimal transport aims to determine the best linear assignment between the elements of two sets under given marginal constraints. To formally describe the proposed approach, we first introduce a cost matrix, C, such that $C_{ij} = c(x_i, y_j)$ is a distance between the vectorized token x_i and token y_j, which are denoted as h_n^S and h_m^T, respectively. Additionally, we introduce two discrete marginal distributions:

$$\Phi = \sum_{n=1}^{N} \phi_n \delta_{h_n^S} \quad ; \quad \Psi = \sum_{m=1}^{M} \psi_m \delta_{h_m^T} \tag{2}$$

where ϕ and ψ are individual weights with respect to each token in $X_{1:N}$ and $Y_{1:M}$ and δ_π is the Dirac function centred on the vector π. The weight vectors are discrete distributions, with their values lying in the simplex (i.e., $\phi_n, \psi_m \geq 0 \; \forall n, m; \sum_{n=1}^{N} \phi_n = \sum_{m=1}^{M} \psi_m = 1$). Hence, optimal transport aims to find a transport matrix, T, achieving the following minimization:

$$O(\Phi, \Psi) = \min_{T \in \Delta(\Phi, \Psi)} \langle T, C \rangle \tag{3}$$

where $\langle \cdot \rangle$ is the Frobenius dot-product and $\Delta(\Phi, \Psi)$ is the set of joint distributions with respective marginals Φ and Ψ. To achieve this minimization, OT matches token pairs of minimum cost from $X_{1:N}$ and $Y_{1:M}$ in a many-to-many manner, respecting their individual weights. Since this convex optimization can be computationally expensive, Cuturi [4] has proposed the Sinkhorn distance, which is an entropy-regularized OT that can be expressed as:

$$O_\epsilon(\Phi, \Psi) = O(\Phi, \Psi) + \epsilon \cdot h(T) \tag{4}$$

where $h(T)$ is the entropy of the transport matrix T and ϵ is a positive regularization coefficient. Note that one of the main benefits of the Sinkhorn distance is that it can be computed efficiently using a dual form:

$$O_\epsilon(\Phi, \Psi) = \langle \Phi, f \rangle + \langle \Psi, g \rangle \tag{5}$$

While the Sinkhorn distance is computationally efficient, it generally leads to a biased solution for a positive ϵ since $O_\epsilon(\Phi, \Phi) \neq 0$, and thus may not perform ideally as a training objective. For this reason, in our work we have chosen to experiment with the recently-proposed *Sinkhorn divergence* [6], which can be formally defined as:

$$\mathcal{L}_{\text{SD}}(\Phi, \Psi) = O_\epsilon(\Phi, \Psi) - \frac{1}{2} O_\epsilon(\Phi, \Phi) - \frac{1}{2} O_\epsilon(\Psi, \Psi) \tag{6}$$

Intuitively, this divergence normalizes the standard Sinkhorn distance by discounting two symmetric terms, $O_\epsilon(\Phi, \Phi)$ and $O_\epsilon(\Psi, \Psi)$, that reflect the intrinsic "hardness" of its arguments, leaving only the alignment contribution in focus. The Sinkhorn divergence in Equation 6, too, can be expressed concisely in dual form by simply subtracting the symmetric terms:

$$\mathcal{L}_{\text{SD}}(\Phi, \Psi) = \langle \Phi, (f - f') \rangle + \langle \Psi, (g - g') \rangle \tag{7}$$

where f', g' are the solutions of the respective symmetric problems.

In our implementation, to cater for the information from the source sentence, we compute the Sinkhorn divergence "contextually" by setting the vectors h_n^S and h_m^T in Equation 2 to the hidden states of the encoder and the decoder. Finally, we compose the seq2seq loss in Equation 1 and the Sinkhorn divergence in Equation 6 into our final training objective, which can be expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{MLE}} + \lambda \cdot \mathcal{L}_{SD} \tag{8}$$

where λ is a hyperparameter that controls the magnitude of the OT component. The objective shows that the proposed approach can be seamlessly incorporated into any contemporary encoder-decoder architecture.

4 Experiments

4.1 Datasets

Machine Translation For this task, we have evaluated our approach on two standard datasets, IWSLT 2014 German↔English (De↔En)[3] and IWSLT 2015 English↔Vietnamese (En↔Vi), and one large-scale dataset (≈ 4M parallel sentences), WMT 2014 English→German (En→De). For IWSLT De↔En and WMT En→De, we perform the same data pre-processing steps as in the Fairseq library[4]. For IWSLT En↔Vi, we use the publicly available dataset[5] with TED tst2012 and tst2013 as validation and test sets, respectively. For the two IWSLT datasets, we have tokenized sentences using the tokenizer and vocabulary from the pretrained mBERT base model, as distributed by Hugging Face.[6]. For WMT

[3] We remark that there are a few misaligned sentence pairs in the official release of this dataset, which end up affecting the test BLEU score. For more details, please refer to https://github.com/pytorch/fairseq/issues/4146. Herein, we report the BLEU scores on the corrected dataset.

[4] https://github.com/pytorch/fairseq/tree/main/examples/translation

[5] https://nlp.stanford.edu/projects/nmt/data/iwslt15.en-vi

[6] https://huggingface.co/bert-base-multilingual-cased

En→De, we have tokenized the dataset using the byte pair encoding (BPE) of Sennrich et al. [20], with 40K subword merge operations. For evaluation, we report the case-sensitive [22] detokenized sacreBLEU score, as suggested by Post et al. [18], and also the recently proposed BERTScore (F_{BERT}) [25] which nicely complements the BLEU score as it is based on embeddings rather than n-grams.

Abstractive Summarization For this task, we have trained our models on the English Gigaword dataset provided by Hugging Face[7]. However, the default dataset contains roughly 190K documents in the validation set, which makes the validation process exceedingly slow. For the sake of efficiency, we have instead used the modified dataset provided by Zhou et al. [27]. For tokenization, we have used the same tokenizer of WMT En→De. For evaluation, we report the ROUGE-1, ROUGE-2 and ROUGE-L on both the original and modified test sets for a comprehensive comparison.

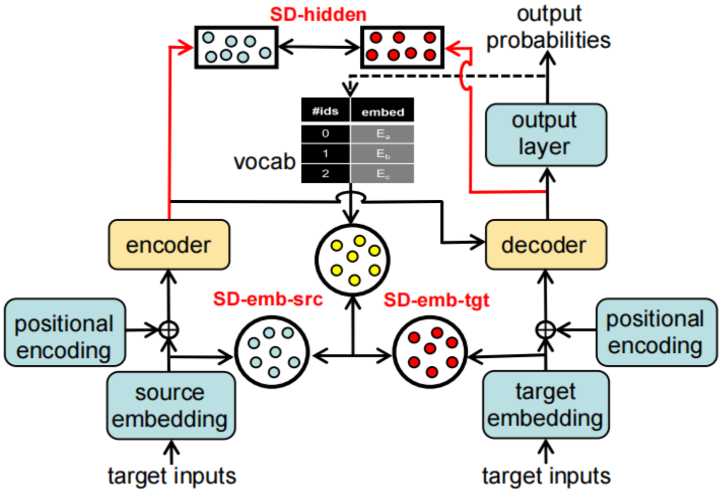

Fig. 1. Illustration of the three different training schemes. The dotted line indicates the path requiring smoothing techniques.

4.2 Models and Training

We have used Fairseq's *transformer_iwslt_de_en* configuration for the two IWSLT tasks and *transformer_wmt_en_de* configuration for the WMT task. For the MLE training, we have used the label-smoothed negative log-likelihood with smoothing parameter 0.1. For our combined training objective, we have explored a wide range of values for the hyperparameter λ in Equation 8 using the IWSLT De→En dataset, and set it to 0.1 for all tasks based on the best performance on the validation set. A sensitivity analysis is presented in Section 4.3. During training, we have batched sentences with a maximum number of 8192

[7] https://huggingface.co/datasets/gigaword

tokens and employed the inverse-sqrt learning rate scheduler and the Adam optimizer with $(\beta_1, \beta_2) = (0.9, 0.98)$. At inference time, we have used beam search decoding with a beam size of 5. For measuring the statistical significance with respect to the baseline model, we have used the paired bootstrap resampling test [9] with 2000 resamples.

The proposed approach minimizes the Sinkhorn divergence between the hidden states of the predicted and source sentences, and we therefore note it as SD-hidden hereafter. In order to single out the respective contributions of the contextual representation and the source text, we have also trained two further models for comparison. In the former, we have computed the same Sinkhorn divergence, but between the word embeddings of the predicted tokens and the source tokens (SD-emb-src), and in the latter, those of the predicted tokens and the target tokens (SD-emb-tgt). It is important to note that, in order to use conventional word embeddings in the two compared models, the predictions from the output layer of the decoder need to be "softened". In our experiments, we have used the probability-averaged embedding as suggested by TFOT [2]. The three different training schemes are summarized in Figure 1.

Table 1. BLEU and F_{BERT} scores for the IWSLT 2014 De↔En and IWSLT 2015 En↔Vi translation tasks. (†) p-value < 0.05. (‡) p-value < 0.01.

Model	**De→En**				**En→De**			
	dev		test		dev		test	
	BLEU	F_{BERT}	BLEU	F_{BERT}	BLEU	F_{BERT}	BLEU	F_{BERT}
Transformer	35.14	67.41	33.44	65.94	29.90	64.27	28.22	63.24
+ SD-emb-src	35.11	67.22	33.39	65.78	29.92	64.41	28.34	63.36
+ SD-emb-tgt	35.27	67.56	33.67†	66.04	29.89	64.44†	28.36	**63.53**‡
+ SD-hidden	**35.59**‡	**67.70**‡	**34.05**‡	**66.31**‡	**30.19**†	**64.66**‡	**28.78**‡	63.46†
Model	**Vi→En**				**En→Vi**			
	dev		test		dev		test	
	BLEU	F_{BERT}	BLEU	F_{BERT}	BLEU	F_{BERT}	BLEU	F_{BERT}
Transformer	24.62	57.96	27.48	61.19	26.59	85.14	29.85	86.63
+ SD-emb-src	24.37	57.90	27.43	61.14	27.10†	85.22	30.28	**86.70**
+ SD-emb-tgt	24.27	57.61	27.57	61.01	26.68	**85.26**†	29.63	86.62
+ SD-hidden	**25.03**	**58.14**	**28.23**†	**61.26**	**27.14**†	**85.26**†	**30.41**†	86.61

4.3 Results and Discussion

Machine Translation We first report the results for the machine translation task over the two IWSLT datasets in Table 1 and the WMT dataset in Table 2. For the two IWSLT translation tasks, both SD-emb models show little or even no improvement compared to the baseline model. However, our SD-hidden model shows consistent improvements over almost all datasets and evaluation metrics, of up to 0.75 pp in BLEU and 0.39 pp in F_{BERT}. Also over the WMT 2014 En→De dataset, our SD-hidden model has performed the best, achieving increases of up to 0.45 pp in BLEU and 0.42 pp in F_{BERT}. Notably, the SD-emb models also show noticeable improvements in this dataset.

For comparison with the literature, we also include other reported sacreBLEU scores over the same datasets.

Table 2. BLEU and F_{BERT} scores for the WMT 2014 En→De translation task. (†) p-value < 0.01. (‡) p-value < 0.001. (⋆) from [21]. (-) not available.

Model	**En→De**			
	dev		test	
	BLEU	F_{BERT}	BLEU	F_{BERT}
Other Reported Results				
Transformer*	-	-	26.5	-
Rel-Transformer*	-	-	26.8	-
Our Implementations				
Transformer	29.99	61.47	26.61	63.27
+ SD-emb-src	$30.22^{\ddagger}$	$61.69^{\ddagger}$	26.97	63.54
+ SD-emb-tgt	$30.20^{\ddagger}$	$61.65^{\ddagger}$	26.95	63.46
+ SD-hidden	**30.3**‡	**61.75**‡	**27.06**†	**63.69**‡

Abstractive Summarization Table 3 shows the results for the abstractive summarization task. For a fair comparison, where available, we report results for both the original and the modified test set. On the original test set, our SD-hidden model performed the best, achieving 0.22 pp, 0.27 pp and 0.23 pp improvements in ROUGE-1, ROUGE-2 and ROUGE-L scores, respectively. Also on the modified test set, it has achieved marked improvements over the baseline of 0.58 pp, 0.32 pp and 0.65 pp in ROUGE-1, ROUGE-2 and ROUGE-L scores, respectively. Yet, in this case, the SD-emb-src approach has slightly outperformed the proposed approach in two metrics. In all cases, all these results confirm the importance of attending to the source information in the training objective.

Table 3. ROUGE-1, ROUGE-2 and ROUGE-L scores for the Gigaword summarization task. (†) p-value < 0.05. (‡) p-value < 0.01. (○) from [24]. (●) from [27]. (◇) from [28]. (-) not available.

Model	**English Gigaword**								
	dev			test			test*		
	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L
Other Reported Results									
Transformer$^{\circ}$	-	-	-	-	-	-	37.57	18.90	34.69
SEASS$^{\bullet}$	-	-	-	46.86	24.58	43.53	36.15	17.54	33.63
SeqCopyNet$^{\diamond}$	-	-	-	47.27	25.07	44.00	35.93	17.51	33.35
Our Implementations									
Transformer	48.00	25.46	44.65	48.35	26.28	44.86	37.90	19.01	35.13
+ SD-emb-src	47.98	25.30	44.51	**49.07**‡	**26.79**	**45.51**‡	37.68	18.64	34.72
+ SD-emb-tgt	47.93	25.31	44.37	48.67	26.39	44.95	37.67	18.59	34.61
+ SD-hidden	**48.42**‡	**25.65**	**45.10**‡	$48.93^{\dagger}$	26.60	**45.51**†	**38.12**	**19.28**	**35.36**

* original test set provided by the Hugging Face library.

Comparison with the Standard OT We have also compared the performance of the Sinkhorn divergence in Equation 6 with the standard OT distance of Equation 4 over the same hidden states. For simplicity, we have limited this comparison to the IWSLT 2014 En→De translation task. The results, reported in Table 4, show that the Sinkhorn divergence has outperformed the standard OT distance in all cases.

Table 4. BLEU and F_{BERT} scores for the standard OT (Sink. Dis. in the table) and the proposed Sinkhorn divergence (Sink. Div. in the table) in the IWSLT 2014 En→De translation task.

Method	**En→De**			
	dev		test	
	BLEU	F_{BERT}	BLEU	F_{BERT}
Transformer	29.90	64.27	28.22	63.24
+ Sink. Dis.	29.81	64.53	28.58	63.40
+ Sink. Div.	**30.19**	**64.66**	**28.78**	**63.46**

Performance Sensitivity to the Value of $\boldsymbol{\lambda}$ Table 5 shows the performance with variable values of λ in the IWSLT 2014 En→De translation task. The key observation is that the results seem reasonably stable, and that values of 0.01 and 0.1 (and, likely, any values in between) have clearly outperformed the baseline (column 0) on the validation set. As a reassuring indication of stability, the same values have also outperformed the baseline on the test set. For convenience, we have used the best value over this validation set (i.e., 0.1) for all tasks.

Table 5. Performance sensitivity to the value of regularization parameter λ for the IWSLT 2014 De→En translation task.

dataset	λ						
	0	0.001	0.01	0.1	0.2	0.5	1
Valid BLEU	35.14	35.22	35.34	**35.59**	35.33	35.13	34.92
Valid F_{BERT}	68.81	68.52	69.03	**69.14**	68.76	68.38	68.39
Test BLEU	33.44	33.38	33.82	**34.05**	33.66	33.61	33.44
Test F_{BERT}	67.39	66.98	67.63	**67.80**	67.58	67.22	67.24

Qualitative Analysis of the Word Alignments To further investigate the proposed approach, in Figure 2 we visualize the optimal transport matrices between the word embeddings of the source and the reference obtained with the different models for a sample from the IWSLT 2014 De→En dataset. For the baseline model, most words are wrongly aligned, especially for the first few tokens (e.g., "also" means "so" in German, "es" means 'it' etc). This shows that the internal attention mechanism of the transformer is not particularly effective at aligning word embeddings. Conversely, the proposed model shows a remarkable performance, even when compared to the two SD-emb models that directly seek to align word embeddings during training. SD-emb-src has the second-best performance, yet it has still failed to correctly align the first three tokens. We have also observed that this behaviour has been even more pronounced in the case of long sentences. For these sentences, our SD-hidden model has displayed a consistent ability to align along the diagonal axis, which is correct in first approximation, while all the other models have predominantly reported very scattered alignments.

It is also noteworthy that the best alignments between word embeddings have not been obtained by the SD-emb objectives that explicitly optimize them. We speculate that a reason for this may be the impact of subword tokenization. For instance, the word *circle* in Figure 2 is an intact token in the English corpus while it has been tokenized into two subwords, *kr* and *##eis*, in the German

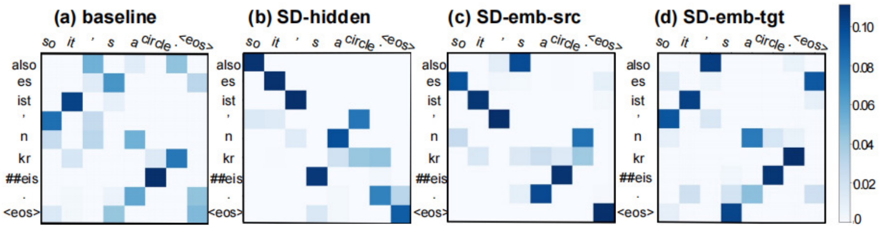

Fig. 2. OT matrices of a sample from the IWSLT 2014 De→En translation task.

corpus. This may somehow break the "equilibrium" in the alignments, as two tokens need to match only one. Since subwords are given the same weight as intact tokens, the unmatched weight may be forcefully assigned to other tokens. We assume that this may be one of the reasons why optimizing the transport directly between word embeddings may lead to poorer alignments. Conversely, optimizing the transport over the hidden states of the transformer may afford more degrees of freedom to mollify this behavior. We leave further investigation to future work.

Table 6. Examples of generated text for the IWSLT 2014 De→En translation task and English Gigaword Summarization task. This table uses color coding to highlight some correct and incorrect phrases. Red: incorrect phrases; Green: correct phrases.

	Translation Examples
Reference	he just looked up at the sky, and he said, "excuse me, can you not see that i'm driving?"
Baseline	he just looked up in heaven and said, you can't see i'm driving? "
SD-emb-src	he just looked up in heaven and said, "so, really, can you see that i'm driving?"
SD-emb-tgt	he just looked up into heaven and said, "forgive me, can't you see i'm driving cars?"
SD-hidden	he just looked up into the sky and said, "excuse me, can't you see that m driving?"
Reference	now these decisions vary in the number of choices that they offer per decision.
Baseline	now these decisions are different in the number of choices that you make when you make choices.
SD-emb-src	now these choices are different from the number of choices that they offer per choice.
SD-emb-tgt	now these choices are different from the number of choices they make.
SD-hidden	now these decisions are different in the number of choices they offer per decision.
	Summarization Examples
Reference	credit agricole announces 1.1-billion-euro bid for greek bank emporiki
Baseline	credit agricole launches offer to buy rest of greek bank
SD-emb-src	credit agricole bids for greek bank
SD-emb-tgt	credit agricole launches 1.1-bln-euro offer for greek bank
SD-hidden	credit agricole bids 1.1 bln euros for emporiki bank
Reference	palestinian official urges arabs to invest in jerusalem
Baseline	palestinian official calls for holy war for east jerusalem
SD-emb-src	palestinian official calls for arab investment in east jerusalem
SD-emb-tgt	palestinian official calls for arab investment in east jerusalem
SD-hidden	palestinian official urges arabs to invest in east jerusalem

Examples of Generated Text In Table 6, we show a few examples from the IWSLT 2014 De→En translation task and the English Gigaword summarization task. Overall, it is easy to appreciate the higher quality of the generated sentences provided by the proposed approach for these samples. In the first translation example, the proposed approach has been the only one that was able to retrieve “sky” (metaphorical) instead of “heaven” (metaphysical) and the “excuse me,” opener. In the second translation example, the proposed approach has been the only one to correctly nuance “different in” and “per decision”. Also in the summarization examples, the proposed model seems to have been the most faithful to the reference. For example, in the second summarization example, the MLE baseline has returned a major mistake by predicting the incorrect phrase “calls for holy war”.

5 Conclusion

In this work, we have proposed a novel training objective for NLG tasks that minimizes the Sinkhorn divergence between the contextual representations of the predictions and the source text. The proposed objective shares the computational efficiency of the well-known Sinkhorn distance and is, in principle, applicable to any of the seq2seq models in common use. The experimental results over various translation and summarization datasets have shown that the proposed approach has been able to achieve statistically-significant improvements over our MLE baseline and two, alternative OT objectives. A qualitative analysis of selected samples has shown that the proposed approach has led to word embeddings that can more effectively align the source and target, even over those of other OT-trained models which are explicitly trained to align word embeddings. In addition, a few examples of generated text have shown that the attention devoted to the source does not come at a price of fluency and adequacy of the generated text.

Acknowledgements The first author is funded by the China Scholarship Council (CSC) from the Ministry of Education of the P. R. of China.

References

1. Alqahtani, S., Lalwani, G., Zhang, Y., Romeo, S., Mansour, S.: Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In: EMNLP (2021)
2. Chen, L., Zhang, Y., Zhang, R., Tao, C., Gan, Z., Zhang, H., Li, B., Shen, D., Chen, C., Carin, L.: Improving sequence-to-sequence learning via optimal transport. In: ICLR. pp. 1–16 (2019)
3. Chen, Y.C., Gan, Z., Cheng, Y., Liu, J., Liu, J.: Distilling knowledge learned in bert for text generation. In: ACL (2020)
4. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: NIPS. pp. 2292–2300 (2013)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)

6. Feydy, J., Séjourné, T., Vialard, F.X., Amari, S., Trouvé, A., Peyré, G.: Interpolating between optimal transport and mmd using sinkhorn divergences. In: AISTATS (2019)
7. Garg, S., Peitz, S., Nallasamy, U., Paulik, M.: Jointly learning to align and translate with transformer models. In: EMNLP (2019)
8. Jauregi Unanue, I., Parnell, J., Piccardi, M.: Berttune: Fine-tuning neural machine translation with bertscore. In: ACL/IJCNLP (2021)
9. Koehn, P.: Statistical significance tests for machine translation evaluation. In: EMNLP (2004)
10. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: ICML (2015)
11. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. ArXiv **abs/1910.13461** (2020)
12. Li, C., Li, J., Wang, G., Fu, H., Lin, Y.C., Chen, L., Zhang, Y., Tao, C., Zhang, R., Wang, W., Shen, D., Yang, Q., Carin, L.: Improving text generation with student-forcing optimal transport. EMNLP pp. 9144–9156 (2020)
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: ACL 2004 (2004)
14. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: EMNLP (2015)
15. Nguyen, T., Luu, A.T.: Improving neural cross-lingual summarization via employing optimal transport distance for knowledge distillation. ArXiv **abs/2112.03473** (2021)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
17. Parnell, J., Unanue, I.J., Piccardi, M.: A multi-document coverage reward for relaxed multi-document summarization. In: ACL (2022)
18. Post, M.: A call for clarity in reporting bleu scores. In: WMT (2018)
19. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks. CoRR **abs/1511.06732** (2016)
20. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. ArXiv **abs/1508.07909** (2016)
21. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: NAACL (2018)
22. Shi, X., Huang, H., Jian, P., Tang, Y.K.: Case-sensitive neural machine translation. Advances in Knowledge Discovery and Data Mining **12084**, 662 – 674 (2020)
23. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS (2014)
24. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. ArXiv **abs/1706.03762** (2017)
25. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTscore: Evaluating text generation with BERT. ArXiv **abs/1904.09675** (2020)
26. Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C.M., Eger, S.: Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. ArXiv **abs/1909.02622** (2019)
27. Zhou, Q., Yang, N., Wei, F., Zhou, M.: Selective encoding for abstractive sentence summarization. In: ACL (2017)
28. Zhou, Q., Yang, N., Wei, F., Zhou, M.: Sequential copying networks. In: AAAI (2018)