Improving Automatic Question Generation Using Wasserstein Distance

Kousshik Raj M (17CS30022)
Supervisor:- Dr. Pawan Goyal

TABLE OF CONTENTS

1

Introduction

What is this about?

NLP

Concerns with the interaction of machine and humans in natural language.

- Symbolic
- Statistical
- Neural

AQG

Generating questions based on given contents. Answers should be present in the contents.

- Text
- Images
- Tables

OBJECTIVE

An algorithm that can be integrated with any generic neural network model for AQG from text and enhance its performance

Look at AQG through the lens of CDA and leverage the recent advances in Optimal Transport in our work

MOTIVATION

 **DATASET**	Generating a large-scale corpus of question - answer pairs of acceptable quality.
 **EDUCATION**	Generating quality questions with no bias and repetition to evaluate student performance.
 **CUSTOMIZATION**	Customizing content into question - answer pairs, for chat boxes and QA systems.

2

Related Works

Question Generation

- *Kalady et al., Ali et al.*, proposed simple rule based approaches involving *wh-movements*.
- For determining the type of question, named entity recognizer, or semantic role labellers were used (*Chen et al.*)
- *Curto et al.,* classified questions based on syntactic structure, question prefix, and answer category, and patterns are learnt for each class separately.
- *Serban et al.,* presented a huge scale QA dataset generated by applying a neural network on the knowledge base, Freebase.
- *Zhou et al.,* employed an attention based encoder-decoder model to generate questions from text

Cross Domain Alignment

- *Yuan et al.*, aimed to identify and optimise the semantic similarities between images and texts in a weakly supervised setup using OT.
- *Yu et al.,* aim to simulate a soft alignment between domains by learning co-attention parameters to infer the relation between them.
- *Chen et al.* formulated the CDA problem as a graph matching problem, and proposed a new innovative framework, Graph Optimal Transport

Wasserstein Distance

- *Xie et al.,* presented a fast proximal gradient method for approximating WD.
- *Rubner et al.,* used WD to model the color distribution structure in image search.
- WD is used as a loss in majority of GANs, to avoid the mode issue collapse (*Goodfellow et al., Salimans et al., Mroueh et al.)*.

3

Background

Traditional Distances

$X \sim P$, $Y \sim Q$. How much difference is there between P and Q ?

Proposed traditional measures:-

- **Total Variation :** $\frac{1}{2}\int |p-q|$
- **Hellinger :** $\sqrt{\int (\sqrt{p}-\sqrt{q})^2}$
- **L_2 :** $\int (p-q)^2$
- **X^2 :** $\int \frac{p-q)^2}{q}$

Here **p** and **q** are probability densities of *P* and *Q*.

Disadvantages

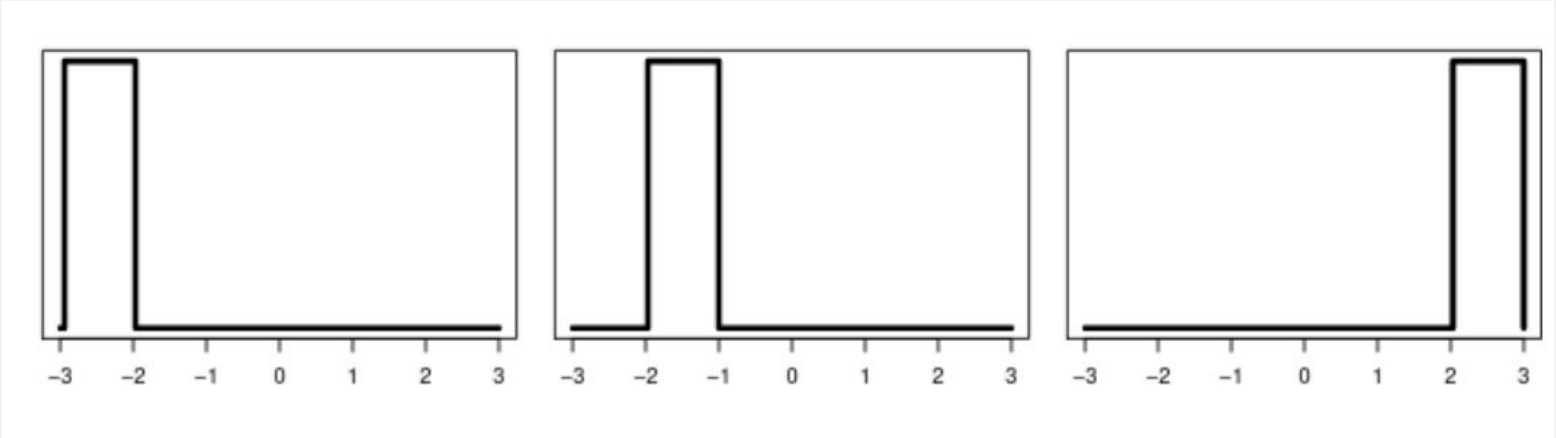
- A proper measure cannot be evaluated when both are not of the same type. Consider P uniform on [0, 1], and Q uniform on {0, 1/N, 2/N, …, 1}. Total variation gives 1, but WD gives 1/N .

- Other than computing the distance, WD also gives us a map telling the optimal way to change P into Q.

- Being sensitive to small disturbances is not beneficial, which is the case for the above distances. WD addresses this, as it is insensitive to small wiggles.

Disadvantages (contd.)

- Normal distances do not take into consideration the underlying geometry of the space. WD is much better in this regard.



Optimal Transport

- Suppose $T : R^d \rightarrow R^d$. Then the distribution of $T(X)$ is called the push forward of P ($T_{\#}P$).

$$T_{\#}P(A) = P(\{y : T(y) \in A\}) = P(T^{-1}(A)).$$

- Optimal Transport Distance:- $\inf_{T} \int ||y - T(y)||^p dP(y),$

  where $T_{\#}P = Q$. Best way to convert P to Q.

- The minimizer T^*, is known as the *optimal transport map*.

Problem?

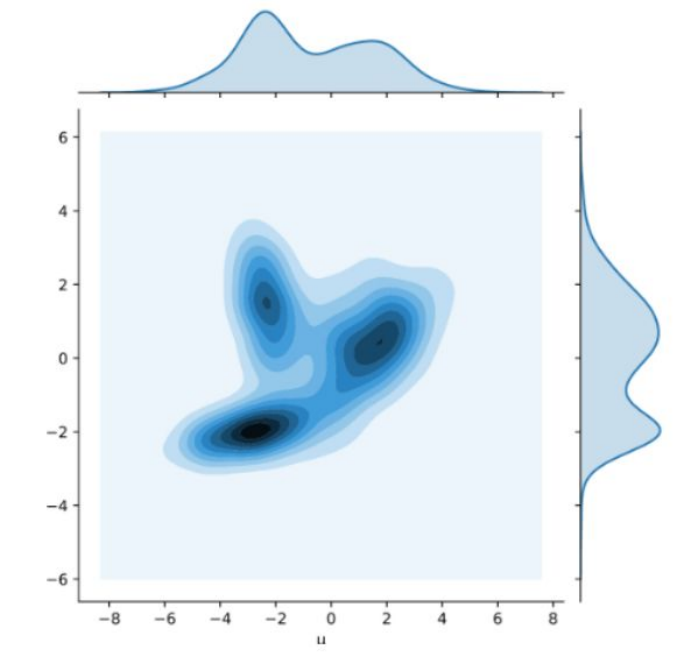
- A map T such that $T_{\#}P = Q$ might not even exist. Why?
- Consider $P = \delta_0$ and $Q = \delta_{-1} / 2 + \delta_1 / 2$. All mass is concentrated in one position in P, whereas there are 2 positions in Q.
- Solution ? Kantorovich modified the current formulation to allow a mass at a particular point to split and move to more than one position.

Wasserstein Distance

Let, J be a joint distribution for (X, Y), marginals being P and Q.

$$W_p(P,Q) = \left(\inf_{J \in \mathcal{J}(P,Q)} \int ||u - v||^p dJ(u,v) \right)^{\frac{1}{p}}$$

- The minimizer J^*, is known as the *optimal coupling*.
- If $p = 1$, commonly known as Earth Mover Distance.



- A joint distribution with marginals shown at the edges.
- There are a lot of such distributions
- Darker color implies higher probability

4

Methodology

Formulation as CDA

Currently,

D_1 ~ input text domain, D_2 ~ target question domain

Let, $A = \{a_1, a_2, \ldots, a_n\}$, $B = \{b_1, b_2, \ldots, b_m\}$ $a_i \in D_1$, $b_j \in D_2$

$$X, Y = f_\theta(A, B)$$ [f_θ represents the deep neural network]

Where, $X = \{x_1, x_2, \ldots, x_n\}$, $Y = \{y_1, y_2, \ldots, y_m\}$

- **A, B** are source and target features
- **X, Y** are contextualized representation of **A, B**

$$L(\theta) = L_{obj}(X, Y)$$ [L is the final training objective]

Formulation as CDA

- But, $L_{obj}(\,.\,)$ only acts as a supervision signal that governs the training of model parameters.
- There is no explicit signal in the objective function that encourages alignment between these two domains.
- To resolve the problem, an additional term corresponding to CDA loss is introduced in the objective function

$$L(\boldsymbol{\theta}) = L_{obj}(\mathbf{X}, \mathbf{Y}) + \alpha \,.\, L_{CDA}(\mathbf{X}, \mathbf{Y})$$

Calculating L_{CDA}

$$\mathcal{L}_{CDA} = W_1(\mu, \nu) = \min_{\mathbf{T} \in \prod(\mathbf{u}, \mathbf{v})} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{T}_{ij} \cdot c(\mathbf{x}_i, \mathbf{y}_j)$$

- **T** is a *n* x *m* matrix representing the mass function of the joint discrete probability distribution of (***u***, ***v***) with marginals being uniform distribution.
- $c(\mathbf{x}_i, \mathbf{y}_j)$ measures the distance between the two embeddings. We use cosine distance

$$c(\mathbf{x}_i, \mathbf{y}_j) = 1 - \frac{\mathbf{x}_i^\top \mathbf{y}_j}{||\mathbf{x}_i||_2 ||\mathbf{y}_j||_2}$$

- Now, L_{CDA} gives an explicit measure of the alignment of the two representations

Algorithm

Input: $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n}$, $\mathbf{Y} = \{\mathbf{y}_j\}_{j=1}^{m}$, D, K

$\boldsymbol{\sigma} = \frac{1}{n}\mathbf{1}_n$

$\mathbf{T}^{(1)} = \frac{1}{nm}\mathbf{1}_n\mathbf{1}_m^\top$

$\mathbf{C}_{ij} = 1 - \frac{\mathbf{x}_i^\top \mathbf{y}_j}{||\mathbf{x}_i|| \cdot ||\mathbf{y}_j||}$

$\mathbf{A}_{ij} = e^{-\mathbf{C}_{ij}}$

for $t = 1, 2, ..., D$ **do**

  $\mathbf{Q} = \mathbf{A} \odot \mathbf{T}^{(t)}$ ▷ $\odot$ is Hadamard product

  for $k = 1, 2, ..., K$ **do**

    $\boldsymbol{\delta} = \frac{1}{n\mathbf{Q}\boldsymbol{\sigma}}$

    $\boldsymbol{\sigma} = \frac{1}{n\mathbf{Q}^\top\boldsymbol{\delta}}$

  $\mathbf{T}^{(t+1)} = diag(\boldsymbol{\delta})\mathbf{Q}diag(\boldsymbol{\sigma})$

$W_1 = \langle \mathbf{C}^\top, \mathbf{T} \rangle$ ▷ $\langle \cdot, \cdot \rangle$ is Frobenius inner-product

Return W_1

Algorithm proposed by Xie et al.

Uses Proximal gradient Method to estimate WD.

D, K, determine the time - accuracy trade off.

C is the cosine distance matrix.

T is the optimal transport plan.

5

Experiments

Model Used

- We will be using the NQG - an attention based encoder decoder model proposed by *Zhou et al*.
- It has an encoder, that takes in a sentence and several handcrafted features like answer features, POS and NER tags, case, etc.
- The decoder takes in the previous target word embeddings, encoder hidden state, and calculates context vector, through which it predicts the next word in sequence.
- Also, has copy mechanism to accommodate for rare words.

Experiment

- We first run the NQG model as it is, and test its performance.
- Then, we integrate our proposed methodology with the NQG framework and again evaluate performance.
- Compare the two scenarios.
- Experiment is done for two different language - English, Bengali.

Dataset

- For English, SQuAD dataset was used. Extracted sentence-question-answer triples. Over 80k samples.
- Used Stanford CoreNLP for the POS and NER tagging.
- For Bengali, TyDiQA dataset was used. Much less samples - around 2k.
- Ignored POS and NER tags, as no way to generate them for Bengali.

Results

- Using BLEU-4 score for evaluating the performance.

Language	NQG	NQG with WD
English	12.97	13.52
Bengali	9.15	9.4

- A significant improvement in English.
- Effect not prominent in Bengali because of low data available for training.
- Base score in Bengali lower than English, as no lexical features were inputted

6

Conclusion

Conclusion

- Started with changing the perspective and looking at AQG as CDA.
- Added an explicit signal to guide the alignment of representation of two domains using Wasserstein Distance, acting as a regularizer.
- Used a fast proximal point optimisation to approximate WD.
- Experiments showed an enhancement in the performance of model after integrating our proposed methodology.

Future Work

- Carry out more experiments on different models and languages.
- Integrate the Optimal Transport map we get while calculating WD into the methodology.
- Currently, only success in generating questions from a sentence. Extend to multi-sentence questions.
- Extend the scope to Visual Question Generation as our methodology doesn’t inherently restrict the source to texts.

References

- Saidalavi Kalady, Ajeesh Illikottil, and Rajarshi das. Natural language question generation using syntax and keywords. I*n Proceedings of QG2010: The Third Workshop on Question Generation*, volume 2, pages 5–14, 2010.
- H. Ali, Y. Chali, and S. A. Hasan. Automation of Question Generation from Sentences. I*n Proceedings of QG2010: The Third Workshop on Question Generation*, Pittsburgh, USA, 2010.
- Wei Chen and Gregory Aist. Generating questions automatically from informational text. I*n Proceedings of the 2nd Workshop on Question Generation (AIED2009)*, pages 17–24, 2009.
- Sérgio Curto, Ana Mendes, and Luisa Coheur. Question generation based on lexicosyntactic patterns learned from the web. *Dislogue and Discourse*, 20(2):147–175, 2012
- Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 588–598, Berlin, Germany, August 2016. Association for Computational Linguistics
- Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. Neural question generation from text: A preliminary study. *CoRR*, abs/1704.01792, 2017.

References

- Siyang Yuan, Ke Bai, Liqun Chen, Yizhe Zhang, Chenyang Tao, Chunyuan Li, Guoyin Wang, Ricardo Henao, and Lawrence Carin. Weakly supervised cross-domain alignment with optimal transport. *arXiv* 2008.06597, 2020.
- Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. *CoRR*, abs/1906.10770, 201
- Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. *arXiv* 2006.14744, 2020
- Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. A metric for distributions with applications to image databases. *In Proceedings of the Sixth International Conference on Computer Vision, ICCV* '98, page 59, USA, 1998. IEEE Computer Society
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. I*n Proceedings of the 27th International Conference on Neural Information Processing Systems* - Volume 2, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press.
- Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact Wasserstein distance. *arXiv* 1802.04307, 2019.

THANK YOU

Questions?