Evaluating Natural Language Generation via Unbalanced Optimal Transport

Yimeng Chen^{1,3*} , **Yanyan Lan**^{1,2†} , **Ruibin Xiong**^{1,2} , **Liang Pang**^{1,2} ,
Zhiming Ma^{1,3} and **Xueqi Cheng**^{1,2}

^{1}University of Chinese Academy of Sciences
^{2}CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, CAS
^{3}Academy of Mathematics and Systems Science, CAS

Abstract

Embedding-based evaluation measures have shown promising improvements on the correlation with human judgments in natural language generation. In these measures, various intrinsic metrics are used in the computation, including generalized precision, recall, F-score and the earth mover's distance. However, the relations between these metrics are unclear, making it difficult to determine which measure to use in real applications. In this paper, we provide an in-depth study on the relations between these metrics. Inspired by the optimal transportation theory, we prove that these metrics correspond to the optimal transport problem with different hard marginal constraints. However, these hard marginal constraints may cause the problem of incomplete and noisy matching in the evaluation process. Therefore we propose a family of new evaluation metrics, namely Lazy Earth Mover's Distances, based on the more general unbalanced optimal transport problem. Experimental results on WMT18 and WMT19 show that our proposed metrics have the ability to produce more consistent evaluation results with human judgements, as compared with existing intrinsic metrics.

1 Introduction

Natural language generation (NLG) has become a hot topic in the area of natural language processing, with a wide range of applications in many tasks, such as image captioning [Xu *et al.*, 2015], machine translation [Bahdanau *et al.*, 2014] and dialogue generation [Li *et al.*, 2016]. To evaluate the performance of NLG methods, qualitative evaluation measures such as BLEU [Papineni *et al.*, 2002], ROUGE [Lin, 2004] and METEOR [Banerjee and Lavie, 2005], are widely used to replace the costly human judgements. However, these measures fail to correlate well with human judgements due to their limitations in considering the semantic meanings of words or phrases, as shown in [Novikova *et al.*, 2017].

Recently, embedding-based evaluation measures have been proposed to tackle with this problem. Typical examples include BERTScore [Zhang *et al.*, 2020], YiSi-1 [Lo, 2019], WMD [Kusner *et al.*, 2015], WMDo [Chow *et al.*, 2019] and MoverScore [Zhao *et al.*, 2019]. They first compute the semantic similarity between word representations, produced by word embedding models such as word2vec [Mikolov *et al.*, 2013] or contextual embedding models such as BERT [Devlin *et al.*, 2018]. The final score is then given by different intrinsic metrics, such as generalized precision, recall and F-score used in BERTScore, and the earth mover's distance (EMD) used in WMD, WMDo and MoverScore.

This paper focus on study the intrinsic metrics. The motivation comes from both empirical and theoretical perspectives. Empirically, it is unclear which intrinsic metric is the best. For example, previous studies have shown that the three different BERTScore versions perform differently on different data sets [Zhang *et al.*, 2020]. Theoretically, existing work on the relations of these intrinsic metrics are superficial. For example, Zhao *et al.* [2019] classifies existing measures to two categories, where EMD is classified as the optimal matching (soft alignment), and generalized precision and recall are classified as the greedy matching (hard alignment). Generalized precision and recall are then transformed to a quasi (non-optimized) EMD form. To the best of our knowledge, an in-depth theoretical study on the relations of these intrinsic metrics is missing.

Inspired by the fact that EMD is a special optimal transport distance [Peyré *et al.*, 2019], we conduct our study from the perspective of optimal transport theory. We theoretically prove that generalized precision and recall correspond to the optimal transport problem with unilateral hard marginal constraint, respectively; while EMD corresponds to the optimal transport problem with bilateral hard marginal constraints. Further considering the fact that F-score is the combination of precision and recall, the relation of different intrinsic metrics is clear. That is, they are the optimal transport problem with different hard marginal constraints. However, these hard constraints may cause serious problems in the NLG evaluation. When the candidate and reference sentences contain paraphrases of different length, only part of the paraphrasing words may be matched under the hard constraints, named *incomplete matching problem*. Besides, some words may be matched to less related ones, instead of their semantically

[*] Work done during the internship at ICT, CAS.

[†] Corresponding author.

close neighbors, named *noisy matching problem*.

To tackle these problems, we propose a family of new intrinsic metrics named Lazy Earth Mover's Distances (Lazy-EMD$_{\lambda_c,\lambda_r}$), induced by the more general unbalanced optimal transport problem with parameters λ_c, λ_r. Existing intrinsic metrics can be viewed as special extreme cases in our framework. Specifically, general precision, recall and EMD corresponds to Lazy-EMD$_{\infty,0}$, Lazy-EMD$_{0,\infty}$ and Lazy-EMD$_{\infty,\infty}$, respectively. Furthermore, we prove that Lazy-EMD has the ability to significantly alleviate the incomplete and noisy matching problem, by replacing the hard marginal constraints to the soft ones. We conduct extensive experiments on the large scale metric evaluation datasets provided by WMT18 and WMT19 [Ma *et al.*, 2018; Ma *et al.*, 2019]. Experimental results show that our proposed metrics produce more consistent results with human judgements, as compared with existing embedding-based evaluation measures.

Our contributions are summarised as follows:

- We theoretically prove the relations of different intrinsic metrics, i.e., the optimal transport problem with different hard marginal constraints.
- We propose a family of new intrinsic metrics based on the more general unbalanced optimal transport problem, which significantly alleviate the incomplete and noisy matching problems of the existing measures.
- Extensive experiments on WMT18 and WMT19 show that our metric outperforms previous ones by producing more consistent results with human judgements.

2 Background

In this section, we introduce some details of the existing embedding-based evaluation measures for natural language generation. Since both BERTScore and YiSi-1 use generalized precision, recall and F-score as the intrinsic metrics, while WMD, WMDo and MoverScore all use earth mover's distance as the intrinsic metric, this section is separated into two subsections according to different intrinsic metrics.

First we give some notations. Suppose a sentence is represented as a triple $(X, \mathbf{X}, \mathbf{w})$, where $X = (x_1, \cdots, x_k)$ is a sequence of tokens, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_k)$ is a sequence of token vectors, $\mathbf{w} = (w_1, \cdots, w_k)$ is a normalized vector with w_i represents the weight assigned to x_i.

2.1 Generalized Precision and Recall

Generalized precision and recall between a candidate sentence $(\hat{X}, \hat{\mathbf{X}}, \hat{\mathbf{w}})$ and a reference sentence $(X, \mathbf{X}, \mathbf{w})$ are defined as follows. They measure the similarity of the two sentences.

Definition 1 (Generalized Precision and Recall)**.**

$$P = \sum_{\hat{\mathbf{x}}_i \in \hat{\mathbf{X}}} \hat{w}_i \max_{\mathbf{x}_j \in \mathbf{X}} S(\hat{\mathbf{x}}_i, \mathbf{x}_j), \tag{1}$$

$$R = \sum_{\mathbf{x}_j \in \mathbf{X}} w_j \max_{\hat{\mathbf{x}}_i \in \hat{\mathbf{X}}} S(\hat{\mathbf{x}}_i, \mathbf{x}_j), \tag{2}$$

where $S(\hat{\mathbf{x}}_i, \mathbf{x}_j)$ measures the similarity of $\hat{\mathbf{x}}_i$ and $\mathbf{x}_j$.

In BERTScore [Zhang *et al.*, 2020], the vector representation $\mathbf{X}$ is generated by the pre-trained model BERT [Devlin *et al.*, 2018]. S is defined as the cosine similarity. The weight vector w_i for each token is either uniform or defined by its inverse document frequency (IDF) computed using the reference sentences. Using generalized precision and recall as the internal metrics, we get two versions of BERTScore, denoted as P_{BERT} and R_{BERT}. F_{BERT} is defined as their harmonic mean.

YiSi-1 is one typical configuration of YiSi [Lo, 2019]. The up-to-date version of YiSi-1 also uses BERT to generate $\mathbf{X}$ for X. The weight w_i is defined by IDF computed using the reference sentences, with a plus-one smoothing in the logarithm. S is defined as the cosine similarity. The final score is defined as a weighted harmonic mean of generalized precision and recall. That is,

$$\text{YiSi-1} = \frac{P \cdot R}{\alpha P + (1 - \alpha) R}, \tag{3}$$

where α is usually set to 0.7 in machine translation evaluation.

2.2 Earth Mover's Distance

The earth mover's distance between the candidate sentence$(\hat{X}, \hat{\mathbf{X}}, \hat{\mathbf{w}})$ and the reference sentence $(X, \mathbf{X}, \mathbf{w})$ is defined as follows, which measures the dissimilarity of the two sentences.

Definition 2 (Earth Mover's Distance)**.**

$$\text{EMD} = \min_{\mathbf{P} \in \mathbb{R}_+^{|\hat{X}| \times |X|}} \langle \mathbf{C}, \mathbf{P} \rangle \tag{4}$$

$$s.t.\ \mathbf{P} \mathbb{1}_{|X|} = \hat{\mathbf{w}}, \mathbf{P}^T \mathbb{1}_{|\hat{X}|} = \mathbf{w}, \tag{5}$$

where $\mathbf{C} \in \mathbb{R}^{|\hat{X}| \times |X|}$, with $C_{i,j}$ stands for the dissimilarity of $\hat{\mathbf{x}}_i$ and $\mathbf{x}_j$. This definition is deduced from the general form of EMD in [Rubner *et al.*, 1998] under the condition $\sum_i \hat{w}_i = \sum_j w_j = 1$.

Word Mover's Distance (WMD) [Kusner *et al.*, 2015] is a typical EMD on two documents, where X is the sequence of content words in the document and $\mathbf{X}$ is the sequence of their word2vec embeddings. $\mathbf{w}$ is a normalized bag-of-words vector. The dissimilarity between two word vectors is measured by the Euclidean distance.

WMD$_o$ [Chow *et al.*, 2019] is an extension of WMD that further incorporates word order. Different from WMD, X is a sequence of both content and stopping words in the sentence, and the dissimilarity is defined based on the cosine similarity. The final score combines WMD with an additional penalty of word order.

MoverScore [Zhao *et al.*, 2019] employs contextual embeddings to replace the word2vec embeddings. Specifically, X and $\mathbf{X}$ are generated by the pre-trained model BERT, and $\mathbf{w}$ is a normalized vector of the IDF weights. Furthermore, the EMD used in MoverScore is an extension of the original one from single words to n-grams.

2.3 Comparisons of Different Metrics

By incorporating embeddings into consideration, these measures have shown promising improvement with human correlation as compared with traditional measures such as BLEU, ROUGE, and METEOR, shown in [Kilickaya *et al.*, 2016; Zhang *et al.*, 2020; Ma *et al.*, 2019]. However, it is not clear which intrinsic metric is the best. For example, previous empirical studies have shown that the three different BERTScore versions perform differently, and alternate as the best on different evaluation datasets [Zhang *et al.*, 2020]. Besides, existing empirical studies on generalized F-score and EMD conducted by BERTScore and MoverScore generate different conclusions on WMT17.

In this case, a thorough theoretical analysis is crucial. However, most existing theoretical results are superficial. In [Zhao *et al.*, 2019], different metrics are related by transforming generalized precision and recall to a quasi non-optimized form of EMD, without further understanding. Some work attempts to separate these metrics to different categories. In [Zhang *et al.*, 2020], the generalized precision and recall, EMD are classified to the greedy and optimal matching, respectively. While in [Zhao *et al.*, 2019], generalized precision and recall, and EMD are viewed as hard and soft alignments, respectively. Though different categories indeed show some connections and differences [Rus and Lintean, 2012], there lacks an in-depth theoretical study on the relations of these metrics. That is exactly the motivation of this paper.

3 Theoretical Analysis on the Relations

Now we show our main theoretical results of the relations between these intrinsic metrics, i.e., they are proven to be the optimal transport problem with different hard marginal constraints. The idea of relating different metrics via optimal transport theory comes from the fact that EMD is an instance of the standard optimal transport distance [Peyré *et al.*, 2019].

3.1 Optimal Transport Problem

Generally, optimal transport problem is to find a transport plan that meets the transport requirements with a minimum cost. Suppose we have two discrete distributions P_X and P_Y supported on $X = (x_1, \cdots, x_n)$ and $Y = (y_1, \cdots, y_m)$, respectively. The amount of mass on these points is given by $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\nu} \in \mathbb{R}^m$. Suppose the cost matrix is $\mathbf{C}$, where $C_{i,j}$ denotes the cost of transporting a unit of mass from point x_i to y_j. We define a coupling matrix $\mathbf{P} \in \mathbb{R}_+^{n\times m}$, where $P_{i,j}$ stands for the amount of mass flowing from the mass at x_i toward y_j. Then the optimal coupling $\mathbf{P}^*$ is the solution of the following standard optimal transport problem.

Definition 3 (Standard Optimal Transport)**.**

$$\min_{\mathbf{P}\in\mathbb{R}_+^{n\times m}} \langle \mathbf{C}, \mathbf{P} \rangle \tag{6}$$

$$\textit{s.t. } \mathbf{P}\mathbb{1}_m = \boldsymbol{\mu}, \mathbf{P}^T\mathbb{1}_n = \boldsymbol{\nu}. \tag{7}$$

From the definition, we can see that a feasible coupling matrix satisfies two hard constraints on the boundaries. Intuitively, if we view $\boldsymbol{\mu}$ as the source and $\boldsymbol{\nu}$ as the target, then $\mathbf{P}\mathbb{1}_m = \boldsymbol{\mu}$ means that all the mass μ_i on the point x_i must be fully transported to Y, and $\mathbf{P}^T\mathbb{1}_n = \boldsymbol{\nu}$ means the total mass transported to y_j must meet its capacity ν_j.

3.2 Relations as Different Constraints

By varying the constraints of Eq. (7), the optimal transport problem bridges generalized precision, recall and EMD.

Firstly, it is straightforward that EMD in Definition 2 corresponds to the optimal transport problem with constraints Eq. (7), by setting $\boldsymbol{\mu} = \hat{\mathbf{w}}$, $\boldsymbol{\nu} = \mathbf{w}$, and $\mathbf{C}$ to be the dissimilarity matrix as in Def. 2. That is to say, EMD could be represented as:

$$\mathrm{EMD} = \langle \mathbf{C}, \mathbf{P}^* \rangle.$$

In fact, generalized precision and recall correspond to the optimal transport problems with each hard marginal constraint in Eq. (7), respectively. Theorem 1 demonstrates how to relate generalized precision to the corresponding optimal transport problem, defined as follows.

$$\min_{\mathbf{P}\in\mathbb{R}_+^{n\times m}} \langle \mathbf{C}, \mathbf{P} \rangle \tag{8}$$

$$\textit{s.t. } \mathbf{P}\mathbb{1}_m = \boldsymbol{\mu}.$$

Theorem 1. *Let* $\boldsymbol{\mu} = \hat{\mathbf{w}}$, $\mathbf{C} = 1 - \mathbf{S}$, *where* $S_{i,j} = S(\hat{\mathbf{x}}_i, \mathbf{x}_j)$ *as defined in Def. 1. Suppose that the solution of the corresponding optimal transport problem (8) is denoted as* $\mathbf{P}_p^*$, *then generalized precision defined in Eq. (1) can be represented as*

$$P = \langle \mathbf{S}, \mathbf{P}_p^* \rangle. \tag{9}$$

Proof. As we discussed in 3.1, the constraint $\mathbf{P}\mathbb{1}_m = \boldsymbol{\mu}$ means all the mass μ_i on the point x_i must be fully transported to Y, while with the constraint $\mathbf{P}^T\mathbb{1}_n = \boldsymbol{\nu}$ removed, there is no requirement on Y. As a result, the optimal plan is transporting all the mass μ_i on the point x_i to the point y_j with the lowest cost $C_{i,j}$. Therefore, the optimal coupling matrix of problem (8) can be written as:

$$(P_p^*)_{i,j} = \begin{cases} \mu_i, & \text{if } j = \text{argmin}_j C_{i,j}. \\ 0, & \text{otherwise.} \end{cases}$$

To prove it, suppose that there exists k, l such that $(P_p^*)_{k,l} > 0$, and $l \neq k^* := \text{argmin}_j C_{k,j}$. Let $\mathbf{P}' = \mathbf{P}_p^*$, except that $P'_{k,l} = 0$, $P'_{k,k^*} = (P_p^*)_{k,l} + (P_p^*)_{k,k^*}$. Then $\mathbf{P}'$ satisfies the constraint, but we have $\langle \mathbf{C}, \mathbf{P}' \rangle < \langle \mathbf{C}, \mathbf{P}_p^* \rangle$, which is a contradiction.

As $\text{argmin}_j C_{i,j} = \text{argmax}_j S_{i,j}$, $\boldsymbol{\mu} = \hat{\mathbf{w}}$, we have

$$(P_p^*)_{i,j} = \begin{cases} \hat{w}_i, & \text{if } j = \text{argmax}_j S_{i,j}. \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The inner product of $\mathbf{S}$ and $\mathbf{P}_p^*$ becomes

$$\langle \mathbf{S}, \mathbf{P}_p^* \rangle = \sum_{i,j} S_{i,j} (P_p^*)_{i,j} = \sum_i \hat{w}_i \max_j S_{i,j}.$$

With $S_{i,j} = S(\hat{\mathbf{x}}_i, \mathbf{x}_j)$, the above formula exactly equals to that of Eq. (1), i.e., generalized precision. □

Similarly, we can prove that generalized recall corresponds to the optimal transport problem with another unilateral hard

		Translations	P	R	F	Lazy-EMD
Example 1	reference	The young man in a slicker.	1	1	1	0
	candidate 1	The boy in a coat.	**0.9560**	0.9419	**0.9489**	0.0533
	candidate 2	The man in a coat.	**0.9609**	0.9408	**0.9507**	0.0553
Example 2	reference	The boy in a coat.	1	1	1	0
	candidate 1	The young man in a slicker.	0.9419	**0.9560**	0.9489	0.0511
	candidate 2	The old man in a slicker.	0.9324	**0.9574**	0.9447	0.0525

Table 1: Evaluation scores under different metrics. Candidate 1 in both examples is better in human judgement. Bold indicates inconsistency.

constraint, i.e., $\mathbf{P}^T\mathbb{1}_n = \boldsymbol{\nu}$, where $\boldsymbol{\nu} = \mathbf{w}$. Let the corresponding optimal coupling matrix denoted as $\mathbf{P}_r^*$, generalized recall can be written as $R = \langle \mathbf{S}, \mathbf{P}_r^* \rangle$.

To conclude, we have proven that generalized precision, recall and EMD correspond to the optimal transport problem with different hard marginal constraints.

3.3 Problems of Hard Constraints

However, these hard constraints may cause some problems in the evaluation process. Consider the coupling value P_{ij}: it can be viewed as an assignment of weights on $\hat{x}_i$ and x_j, which are lexical tokens in the candidate and reference sentences, respectively. When the value is non-zero, we say the corresponding tokens are matched. In this paper, our theoretical analysis shows that the hard constraints may cause incomplete and noisy matching problems.

The incomplete matching problem means that only part of the paraphrasing words are matched, which usually happens when the candidate and reference sentences contain paraphrases of different length. Now we show why the hard constraints result in this phenomenon.

First consider generalized precision, suppose that word $\hat{x}_1$ in the candidate sentence is a paraphrase of words x_1 and x_2 in the reference sentence. By the explicit form of $\mathbf{P}_p^*$ in Eq. (10), $\hat{x}_1$ will only be matched with one of x_1 and x_2, unless $S(\hat{x}_1, x_1)$ equals exactly as $S(\hat{x}_1, x_2)$, which is very rare even though x_1 and x_2 are similar. Similarly, when candidate sentences contain longer paraphrase, the incomplete matching problem will happen for generalized recall.

Table 1 shows two examples of the incomplete matching problem, corresponding to the generalized precision and recall, respectively. In both examples, 'boy' is matched with 'man', instead of 'young man'. Evaluation results given by generalized precision and recall are both contradicted by human judgements. As the harmonic mean of them, F-score sometimes fixes their bias as in Example 2. However in Example 1, F-score still shows inconsistency. We also show the results of our proposed new metric Lazy-EMD, of which the dissimilarity evaluations agree with human judgements.

For EMD, the bilateral hard constraints further limit the capacity of semantic units. As a result, the matching is dependent on the value of $\hat{w}_1, \hat{w}_2$ and w_1. If $w_1 > \min(\hat{w}_1, \hat{w}_2)$, both $\hat{x}_1$ and $\hat{x}_2$ will be matched with x_1. However if $w_1 \leq \min(\hat{w}_1, \hat{w}_2)$, since the optimal $\mathbf{P}^*$ must satisfies $\sum_i P_{i,1}^* = w_1$, x_1 will be only matched with the nearest one between $\hat{x}_1$ and $\hat{x}_2$. That is the incomplete matching problem.

Noisy matching problem means that words are matched with some less relevant tokens, instead of their semantic neighbors. We still consider the above example, with further assumption that the rest of the words in the reference sentence are less relevant with either $\hat{x}_1$ or $\hat{x}_2$ semantically. For EMD, when $w_1 = \hat{w}_1$, $C_{1,1} < C_{2,1}$, coupling $P_{1,1}^* = \hat{w}_1$. By the constraint on $\mathbf{P}^T\mathbb{1}_n$, $\sum_i P_{i,1}^* = \hat{w}_1$, then we have $P_{2,1}^* = 0$. Since the constraint requires $\sum_j P_{2,j}^* = \hat{w}_2$, there exists some k s.t. $P_{2,k}^* > 0$. In other words, instead of x_1, $\hat{x}_2$ will be matched to some unrelated words. That is exactly the noisy matching problem.

4 Lazy Earth Mover's Distances

To tackle the above problems, we propose a family of new intrinsic metrics, namely Lazy Earth Mover's Distances, induced from the more general unbalanced optimal transport problem. Unbalanced optimal transport problem [Peyré *et al.*, 2019] relaxes the hard marginal constraints in standard optimal transport, by incorporating them into the optimization objective as penalties.

Definition 4 (Unbalanced Optimal Transport)**.**

$$\min_{\mathbf{P}\in\mathbb{R}_+^{n\times m}} \langle \mathbf{C}, \mathbf{P} \rangle + \lambda_c \mathrm{KL}(\mathbf{P}\mathbb{1}_m | \boldsymbol{\mu}) + \lambda_r \mathrm{KL}(\mathbf{P}^T\mathbb{1}_n | \boldsymbol{\nu}).$$

In the definition, parameters λ_c and λ_r control how much the corresponding marginal deviation are penalized, measured by K-L divergence [Kullback and Leibler, 1951].

Let $\boldsymbol{\mu} = \hat{\mathbf{w}}$, $\boldsymbol{\nu} = \mathbf{w}$, and $\mathbf{C}$ be the dissimilarity matrix as in Thm. 1. Denote $\mathbf{P^*}_{\lambda_c,\lambda_r}$ as the corresponding optimal coupling matrix of the unbalanced optimal problem with penalty parameters λ_c, λ_r. The Lazy Earth Mover's Distance between sentences is then defined as follows.

$$\text{Lazy-EMD}_{\lambda_c,\lambda_r} = \langle \mathbf{C}, \mathbf{P^*}_{\lambda_c,\lambda_r} \rangle \tag{11}$$

The word 'lazy' is to emphasize the fact that the final transport plan may not meet its original requirement. The transport behavior will be further explained in Section 4.1.

From the above definitions, we can see that standard optimal transport problem is a special case of unbalanced optimal transport problem, with $\lambda_c = \infty$, $\lambda_r = \infty$. Furthermore, $\mathbf{P}_p^*$ and $\mathbf{P}_r^*$ are the solutions of the unbalanced transport problem with $\lambda_c = \infty$, $\lambda_r = 0$, and $\lambda_c = 0$, $\lambda_r = \infty$, respectively. Therefore, the intrinsic metrics generalized precision, recall, and EMD can be viewed as special Lazy-EMDs:

$$\text{EMD} = \text{Lazy-EMD}_{\infty,\infty},$$

$$P = 1 - \text{Lazy-EMD}_{\infty,0}, \;\; R = 1 - \text{Lazy-EMD}_{0,\infty}.$$

	cs-en	**de-en**	**et-en**	**fi-en**	**ru-en**	**tr-en**	**zh-en**
n	5k/5k	78k/20k	57k/32k	16k/10k	10k/22k	9k/1k	33k/29k
SENTBLEU	.233/.389	.415/.620	.285/.414	.154/.355	.228/.330	.145/.261	.178/.311
P_{BERT}	.387/.539*	.541/.713*	.389/.547*	.283/.482*	.345/.418*	.280/.339*	.248/.337
R_{BERT}	.388/**.571***	.546/.727*	.391/**.591***	.304/.561*	.343/.421*	.290/.395*	.255/.368*
F_{BERT}	.404/.559*	.550/.726*	.397/.584*	.296/.538*	.353/**.424***	.292/.389*	.264/.364
YiSi-1	**.406**/.562	**.551/.731**	.396/.589	**.305**/.551	.350/**.424**	**.294**/.401	.262/.370
F_α	.404/.570	.550/.728	**.398/.591**	.296/**.563**	**.354**/.422	.289/.398	**.265**/.370
EMD	.393/.548	.540/.718	.389/.585	.291/.528	.336/.416	.276/.364	.263/.371
Lazy-EMD	.403/.555	.544/.730	.395/**.591**	.301/.562	.350/**.424**	.281/**.404**	.264/**.372**

Table 2: Kendall's correlations of different metrics with segment-level human judgements on WMT18. For each language pair, the left number is the correlation of to-English, and the right is that of from-English. Highest correlation scores for each language pair are highlighted in bold. Numbers with $*$ are slightly different from those in [Zhang *et al.*, 2020].

4.1 Optimal Transport Plan of Lazy-EMD

Intuitively, by replacing hard marginal constraints with soft ones, Lazy-EMDs have the ability to alleviate the incomplete and noisy matching problems. Now we theoretically prove the above claim, by investigating the optimal transport plan of the unbalanced optimal transport problem.

Suppose that the candidate sentence is mapped to $\hat{X} = (\hat{x}_1, \cdots, \hat{x}_n)$, with a weight vector $\hat{\mathbf{w}} = (\hat{w}_1, \cdots, \hat{w}_n)$ representing the mass on each token. For simplicity, we consider the case where the reference sentence contains only one word $X = (x_r)$, with a unit of mass on it. Let the transport cost from x_r to $\hat{x}_i$ be $c_i > 0, \forall i$, and the transport plan be a vector $\mathbf{P} = (p_1, \cdots, p_n)^T$. Then the optimal transport plan can be obtained with closed form shown in the following theorem.

Theorem 2. *The optimal transport plan $\mathbf{P}^*_{\lambda_c,\lambda_r}$ of the unbalanced transport problem with penalty parameters $0 < \lambda_c, \lambda_r < \infty$ satisfies*

$$p_i^* = \exp(-\frac{c_i}{\lambda_c} - \frac{\lambda_r}{\lambda_c}A) \cdot \hat{w}_i, \tag{12}$$

$$A = \log \sum_i p_i^*. \tag{13}$$

From the results, we can see that the optimal transport plan is obtained by reweighting the mass vector $\hat{\mathbf{w}}$. Specifically, the coefficient $\exp(-\frac{c_i}{\lambda_c} - \frac{\lambda_r}{\lambda_c}A)$ is a decreasing function of c_i. If the incomplete matching problem happens for $\hat{x}_i$, $\hat{x}_i$ must be semantically close with x_r, i.e., $S_{i,r}$ is large. Therefore, $c_i = 1 - S_{i,r}$ is small, and the reweighting coefficient of p_i^* will be large. In this way, the incomplete matching problem will be significantly alleviated. Similarly, the noisy matching problem usually happens for less relevant tokens with large costs. Therefore, the reweighting coefficients in the optimal transport plan on these tokens will be small. In this way, the noisy matching problem will be alleviated.

We further analyze the impact of penalty parameters λ_c and λ_r on the optimal coupling $\mathbf{P}^*$. It is clear that p_i^* increases with λ_c and λ_r. So the unilateral and bilateral upper bounds are achieved by the optimal coupling correspondent to generalized precision, recall and EMD. That is,

$$\mathbf{P}^*_{\infty,\infty} = \hat{\mathbf{w}}, \ \mathbf{P}^*_{\infty,0} = \hat{\mathbf{w}}, \tag{14}$$

$$(\mathbf{P}^*_{0,\infty})_j = \delta_i(j), \text{with } i = \text{argmin}_j \, c_i. \tag{15}$$

5 Experiments

This section demonstrates our experimental results on WMT18 [Ma *et al.*, 2018] and WMT19 [Ma *et al.*, 2019]. We evaluate six different intrinsic metrics, i.e. generalized precision, recall, F1, F_α, WMD and Lazy-EMD, by their segment-level correlations with human judgments.

5.1 Datasets

Our experiments are conducted on WMT18 and WMT19, two widely used machine translation datasets for evaluating NLG measures. Specifically, WMT18 contains predictions of 149 translation systems across 14 language pairs, while WMT19 contains predictions of 193 translation systems across 15 language pairs. We follow the WMT18 and WMT19 standard practice and use Kendall rank correlation to evaluate metric quality on the segment-level human judgements, where WMT18 contains the relative ranking result of 327k sentence pairs and WMT19 contains 531k sentence pairs.

5.2 Implementation Details

For a fair comparison, we fix the embeddings and focus on comparing different intrinsic metrics. According to previous studies [Peters *et al.*, 2018; Devlin *et al.*, 2018], the contextual embeddings produced by BERT is usually better than word2vec-based word embeddings for various downstream NLP tasks. So we apply the default setting of BERTScore to other measures in our comparison. The implementation is based on BERTScore v0.2.2.

For EMD computation, the built-in function `ot.emd` in python package POT is used [Flamary and Courty, 2017]. Lazy-EMD is computed with the generalized Sinkhorn scaling algorithm.[1] The regularization parameter in the Sinkhorn-scaling algorithm is set as 0.009. The penalty parameters are set to be different for three data categories, based on the target language of the translation, i.e., English, Chinese and others. For English, the parameter is set to (0.23, 0.31), which is tuned on et-en in WMT18. For Chinese, the parameter is set as (0.018, 0.97), which is tuned on en-zh in WMT19. For other languages, the parameter is set as (0.009, 0.95), which is tuned on en-cs in WMT19.

[1] Code is available at https://github.com/Beastlyprime/lazy_emd

	cs-en	**de-en**	**fi-en**	**gu-en**	**kk-en**	**lt-en**	**ru-en**	**zh-en**
n	-/27k	85k/100k	38k/32k	31k/11k	27k/18k	22k/17k	46k/24k	31k/19k
sentBLEU	-/.367	.056/.248	.233/.396	.188/.465	.377/.392	.262/.334	.125/.469	.323/.270
P_{BERT}	-/.444	.156/.314	.326/.498	.307/.519	.419/.493	.375/.422	.212/.540	.410/.306
R_{BERT}	-/.494	.160/.351	**.346**/.521	.295/.562	.416/**.541**	.367/.449	.216/.577	.427/.352
F_{BERT}	-/.479	.166/.338	.344/.518	.313/.554	**.434**/.532	.375/.448	.223/.572	.430/.347
YiSi-1	-/.486	.165/.345	**.346**/.521	.317/.563	.433/.538	.373/.450	**.225**/.575	**.433**/.353
F_α	-/.495	.165/.351	.344/.522	.314/.563	**.434/.541**	.375/.449	.223/.578	.429/**.357**
EMD	-/.479	.159/.338	.342/.523	**.318**/.561	.432/.539	**.377**/.455	.215/.566	.430/.343
Lazy-EMD	-/**.498**	**.174/.356**	**.346/.526**	**.318/.569**	.431/**.541**	**.377/.466**	.215/**.582**	**.433**/.352

Table 3: Kendall's correlations of different metrics with segment-level human judgements on WMT19, with notations similar to that on WMT18. The results on '**cs-en**' is missing because there is no such data on WMT19.

	cs-en	**de-en**	**fi-en**	**gu-en**	**kk-en**	**lt-en**	**ru-en**	**zh-en**
(λ_c, λ_r)	-/27k	85k/100k	38k/32k	31k/11k	27k/18k	22k/17k	46k/24k	31k/19k
(0.23, 0.31)	-/.487	**.174**/.351	**.346**/.523	**.318**/.562	**.431**/.531	**.377/.471**	**.215**/.579	**.433**/.337
(0.009, 0.95)	-/**.498**	.172/**.356**	.343/**.526**	.292/**.570**	.413/**.541**	.369/.466	.213/**.582**	.427/.351
(0.018, 0.97)	-/.497	.174/.355	.343/**.526**	.293/.569	.415/**.541**	.368/.467	.214/.581	.426/**.352**

Table 4: Influences of different parameters for Lazy-EMD on WMT19.

We compare Lazy-EMD with generalized precision, recall, F1 and EMD. Further considering that Lazy-EMD is a soft EMD which involves additional parameters, we also compare with F_α, where the parameter α is used to balance the effect of precision and recall. F_α can be viewed as an extension of F1 used in YiSi-1 ($\alpha = 0.7$). In our experiments, α is tuned in the same way as we do for Lazy-EMD. Specifically, α is set to 0.48, 0.9, and 0.96 for English, Chinese and other target languages, respectively.

5.3 Experimental Results

Table 2 and Table 3 show our main experimental results. From the results, we can see that all the embedding-based measures outperform the n-gram based evaluation measure sentBLEU [Ma *et al.*, 2018]. However, since WMT18 is relatively not large, all the embedding-based metrics perform comparably. For the larger dataset WMT19, Lazy-EMD outperforms all baselines in all translation scenarios, even better than F_α. Specifically, Lazy-EMD achieves the best correlation on 12 of 15 language pairs, which validates the advantage of replacing hard marginal constraints with soft ones.

Since the results in Table 2 and Table 3 are under parameters specifically tuned for different target languages, we show the performances of Lazy-EMD under the three different parameters on WMT19, to further study the influence of different penalty parameters. The results are demonstrated in Table 4. We can see that optimal parameter choices do differ between languages. However, the bottom two lines show the performance of Lazy-EMD is insensitive to slight variation on the parameters, which is another benefit of the proposed metric. The analysis of the origin of the difference in optimal parameters will be an interesting topic for future research.

6 Conclusions

This paper focuses on studying different intrinsic metrics of existing embedding-based evaluation measures, i.e., generalized precision, recall, F-score and EMD. We theoretically prove that these intrinsic metrics correspond to the optimal transport problem with different hard marginal constraints. To tackle the incomplete and noisy matching problems of these intrinsic metrics, we propose a family of new metrics, namely Lazy-EMD, based on the more general unbalanced optimal transport problem. Extensive experiments on WMT18 and WMT19 show that Lazy-EMD outperform traditional embedding-based measures in terms of consistency with segment-level human judgements.

To the best of our knowledge, this is the first in-depth theoretical study on the relations of different embedding-based NLG evaluation measures. The main novelty lies in the viewpoint of optimal transport theory. In the future, we plan to extend Lazy-EMD to evaluate distance between documents by incorporating document structure into the unbalanced optimal transport problem.

Acknowledgments

This work was supported by Beijing Academy of Artificial Intelligence (BAAI) under Grants No. BAAI2019ZD0306, and BAAI2020ZJ0303, the National Natural Science Foundation of China (NSFC) under Grants No. 61722211, 61773362, 61872338, 61902381, and 61906180, the Youth Innovation Promotion Association CAS under Grants No. 20144310, and 2016102, the National Key RD Program of China under Grants No. 2016QY02D0405, the Lenovo-CAS Joint Lab Youth Scientist Project, and Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202033).

References

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[Chow *et al.*, 2019] Julian Chow, Lucia Specia, and Pranava Swaroop Madhyastha. Wmdo: Fluency-based word mover's distance for machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 494–500, 2019.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Flamary and Courty, 2017] R'emi Flamary and Nicolas Courty. Pot python optimal transport library, 2017.

[Kilickaya *et al.*, 2016] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. *arXiv preprint arXiv:1612.07600*, 2016.

[Kullback and Leibler, 1951] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[Kusner *et al.*, 2015] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966, 2015.

[Li *et al.*, 2016] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[Lo, 2019] Chi-kiu Lo. Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, 2019.

[Ma *et al.*, 2018] Qingsong Ma, Ondřej Bojar, and Yvette Graham. Results of the wmt18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the third conference on machine translation: shared task papers*, pages 671–688, 2018.

[Ma *et al.*, 2019] Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, 2019.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Novikova *et al.*, 2017] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*, 2017.

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

[Peters *et al.*, 2018] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[Peyré *et al.*, 2019] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[Rubner *et al.*, 1998] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 59–66. IEEE, 1998.

[Rus and Lintean, 2012] Vasile Rus and Mihai Lintean. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162. Association for Computational Linguistics, 2012.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[Zhang *et al.*, 2020] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[Zhao *et al.*, 2019] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, 2019.