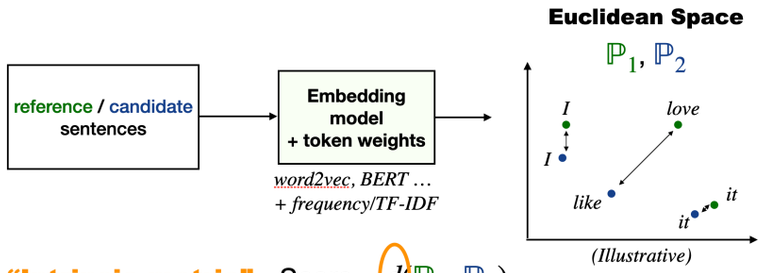
Lazy Earth Mover's Distance is a Better Intrinsic Metric

TRY IT! → https://github.com/Beastlyprime/lazy_emd

MOTIVATION

- Embedding-based evaluation measures have shown promising improvements
- Various intrinsic metrics are used in these measures
- The relations between these metrics are unclear, making it difficult to determine which measure to use in real applications.

INTRINSIC METRICS



"Intrinsic metric" Score = $d(\mathbb{P}_1, \mathbb{P}_2)$

Existing Intrinsic Metrics:

- Generalized precision, recall, F-score (In BERTScore (ICLR, 2019))
- Earth mover's distance (In WMD (ICML, 2014), MoverScore (EMNLP, 2019))

We prove that they correspond to optimal transport plan under different hard constraints

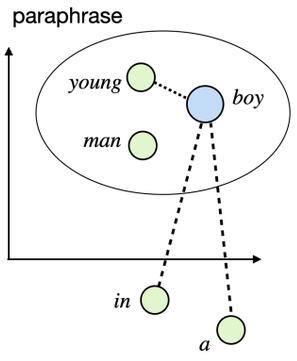
$$\min_{\mathbf{P}\in\mathbb{R}_+^{n\times m}} \langle \mathbf{C}, \mathbf{P}\rangle$$

$$s.t.\ \mathbf{P}\mathbb{1}_m = \boldsymbol{\mu}, \mathbf{P}^T\mathbb{1}_n = \boldsymbol{\nu}. \longrightarrow EMD = \langle C, P^*\rangle$$
$$s.t.\ \mathbf{P}\mathbb{1}_m = \boldsymbol{\mu} \longrightarrow P = \langle S, P_p^*\rangle$$
$$s.t.\ \mathbf{P}^T\mathbb{1}_n = \boldsymbol{\nu}. \longrightarrow R = \langle S, P_r^*\rangle$$

MATCHING PROBLEMS

Reference: The young man in a slicker.

Candidate: The boy in a coat.

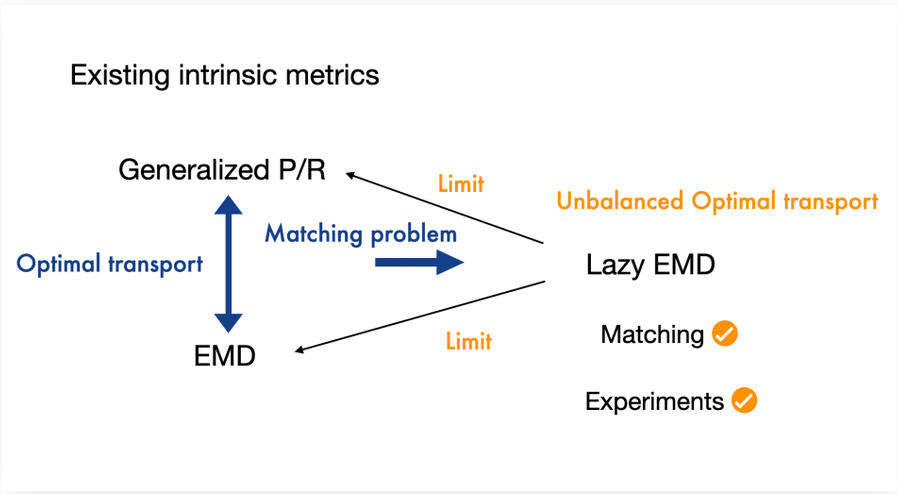


- Incomplete matching: happens when paraphrases in two sentences are partly matched.
- Noisy matching: happens when words are matched to less related ones, instead of their semantic neighbors.

We prove that Generalized Precision, Recall and EMD have the above two problems.

HIGHLIGHTS

- Existing intrinsic metrics can be bridged by optimal transport problem. They correspond to optimal transport plan under different hard constraints on the marginal.
- Existing intrinsic metrics induce incomplete and noisy matching, due to the hard constraints.
- We propose Lazy Earth Mover's Distance, an intrinsic metric induced by optimal transport problem with soft bilateral constraints.
- Theoretically and experimentally, evaluation measure based on Lazy Earth Mover's Distance produces better evaluation results.



Lazy-EMD

Optimal Transport

$$\min_{\mathbf{P}\in\mathbb{R}_+^{n\times m}} \langle \mathbf{C}, \mathbf{P}\rangle$$
$$s.t.\ \mathbf{P}\mathbb{1}_m = \boldsymbol{\mu}, \mathbf{P}^T\mathbb{1}_n = \boldsymbol{\nu}.$$

Unbalanced Optimal Transport

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

soft → $\mathbf{P}^*_{\lambda_c,\lambda_r}$

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

- Lazy-EMD is induced from unbalanced optimal transport problem, which relaxes the hard marginal constraints.
- Lazy-EMD recovers EMD, Generalized P, R at the limits:

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

- Lazy matching alleviates the incomplete and noisy matching problems: (c: distance p: matching weight)

$$p_i^* = \exp\left(-\frac{c_i}{\lambda_c} - \frac{\lambda_r}{\lambda_c}A\right)\cdot w_i \quad c_i \nearrow, p_i^* \searrow$$

RESULTS

Experiments results on WMT19 translation benchmark.

- Evaluation quality is measured by segment level correlations with human judgements.
- Results show Lazy-EMD achieves the best on 12 of 15 language pairs

n	cs-en -/27k	de-en 85k/100k	fi-en 38k/32k	gu-en 31k/11k	kk-en 27k/18k	lt-en 22k/17k	ru-en 46k/24k	zh-en 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

We use different penalty parameters λ_r, λ_c for 3 kinds of different target languages (English, Chinese and others). The following table shows the performances of Lazy-EMD under the three different parameters on WMT19, to further study the influence of different penalty parameters.

(λ_c, λ_r)	cs-en -/27k	de-en 85k/100k	fi-en 38k/32k	gu-en 31k/11k	kk-en 27k/18k	lt-en 22k/17k	ru-en 46k/24k	zh-en 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**

It shows:

- The performance of Lazy-EMD is insensitive to slight variation on the parameters.
- Optimal parameter choices differ between languages.