IJCAI-PRICAI
YOKOHAMA
2020

Evaluating Natural Language Generation via Unbalanced Optimal Transport

Yimeng Chen, Yanyan Lan, Ruibin Xiong, Liang Pang,
Zhiming Ma and Xueqi Cheng

Outline

Part 1 - A Brief Introduction

Part 2 - More Details

Part 1

A Brief Introduction

Part 1 - Outline

- Motivation
- 3 Highlights
  - Bridging by optimal transport
  - Matching problems
  - Lazy Earth Mover's Distance
- Experiment results
- Conclusion

Outline

- Motivation
- 3 Highlights
  - Bridging by optimal transport
  - Matching problems
  - Lazy Earth Mover's Distance
- Experiment results
- Conclusion

Outline

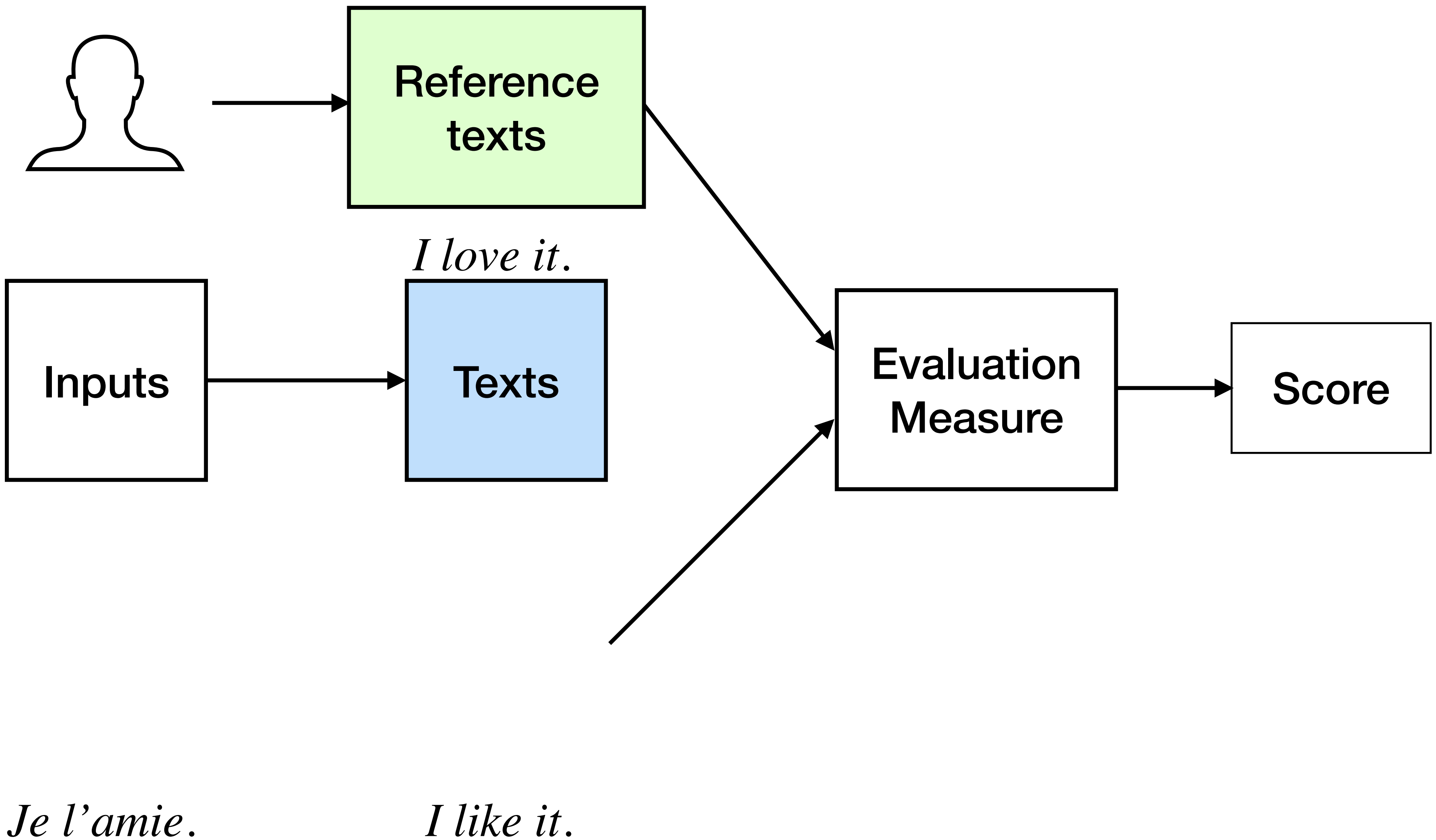
- Motivation
- 3 Highlights
  - Bridging by optimal transport
  - Matching problems
  - Lazy Earth Mover's Distance
- Experiment results
- Conclusion

Code and demo: https://github.com/Beastlyprime/lazy_emd

Motivation

Q: Which intrinsic metric is better for embedding-based NLG evaluation measures?

Natural Language Generation Evaluation

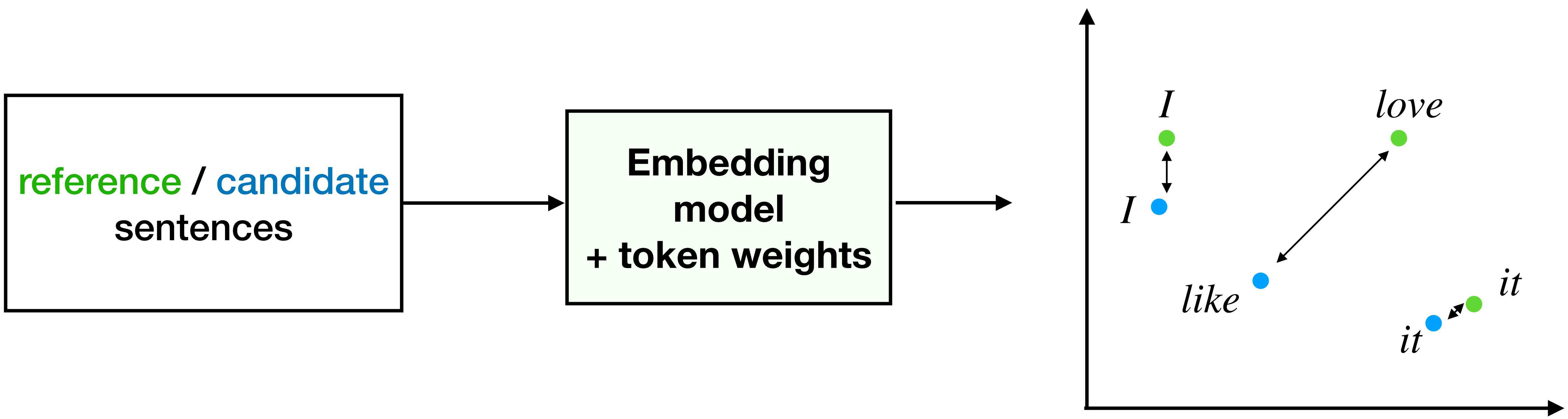


Reference: https://iclr.cc/virtual_2020/poster_SkeHuCVFDr.html

Embedding-Based Measures

Euclidean Space

$\mathbb{P}_1, \mathbb{P}_2$



(Illustrative)

$$\text{Score} = d(\mathbb{P}_1, \mathbb{P}_2)$$

“Intrinsic metric”

Existing intrinsic metrics

Generalized precision/recall

- BERTScore (ICLR 2020)
- YiSi-1 (CMT 2019)

Earth mover's distance

- WMD (ICML 2015)
- WMDo (CMT 2019)
- MoverScore (EMNLP 2019)

? Which is the best? Difference? Relations?

Highlight

1

Bridging by Optimal Transport

Different HARD constraints

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

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

$$s.t.\ \mathbf{P}\mathbb{1}_m = \boldsymbol{\mu} \longrightarrow P = \langle S, P_p^* \rangle$$

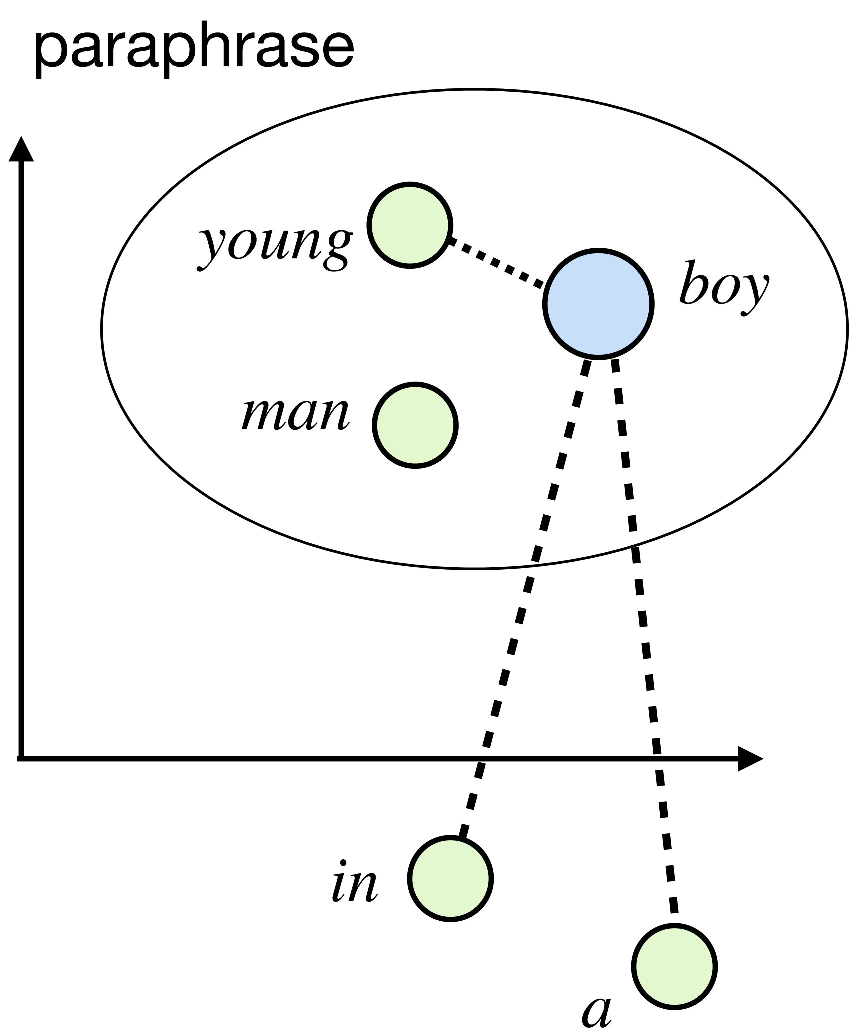
$$s.t.\ \mathbf{P}^T\mathbb{1}_n = \boldsymbol{\nu}. \longrightarrow R = \langle S, P_r^* \rangle$$

Highlight

2

Matching Problems

Existing Metrics Induce BAD match



1. Incomplete matching

2. Noisy matching

HARD Constraints, BAD Match

		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

		Captions	EMD	Lazy-EMD
Example 3	reference	A dog runs in the grass.	0	0
	caption 1	A boy climbs up the tree.	**0.0738**	0.4301
	caption 2	A playful dog is running through the grass.	**0.0881**	0.3104
Example 4	reference	A boy climbs up the tree.	0	0
	caption 1	A dog runs in the grass.	**0.0738**	0.4301
	caption 2	A brave boy is climbing up a tall tree.	**0.0781**	0.3491

Bad match → **inconsistent evaluation**

Highlight

3

Lazy Earth Mover’s Distance

Lazy Earth Mover's Distance

- Unbalanced Optimal Transport Problem

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

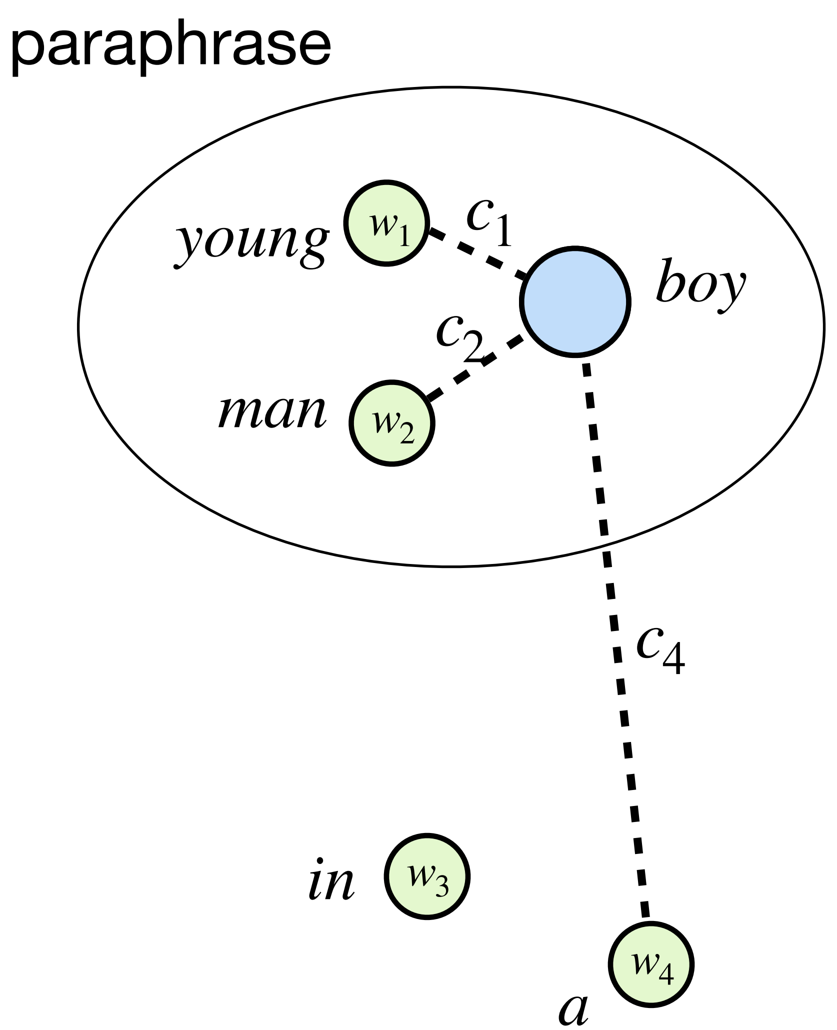
$\downarrow$ $\mathbf{P}^*_{\lambda_c,\lambda_r}$

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

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

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

Lazy matching $\mathbf{P}^*_{\lambda_c,\lambda_r}$



Matching weight

$c_i \nearrow, p_i^* \searrow$

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

That alleviate the incomplete and noisy matching problems!

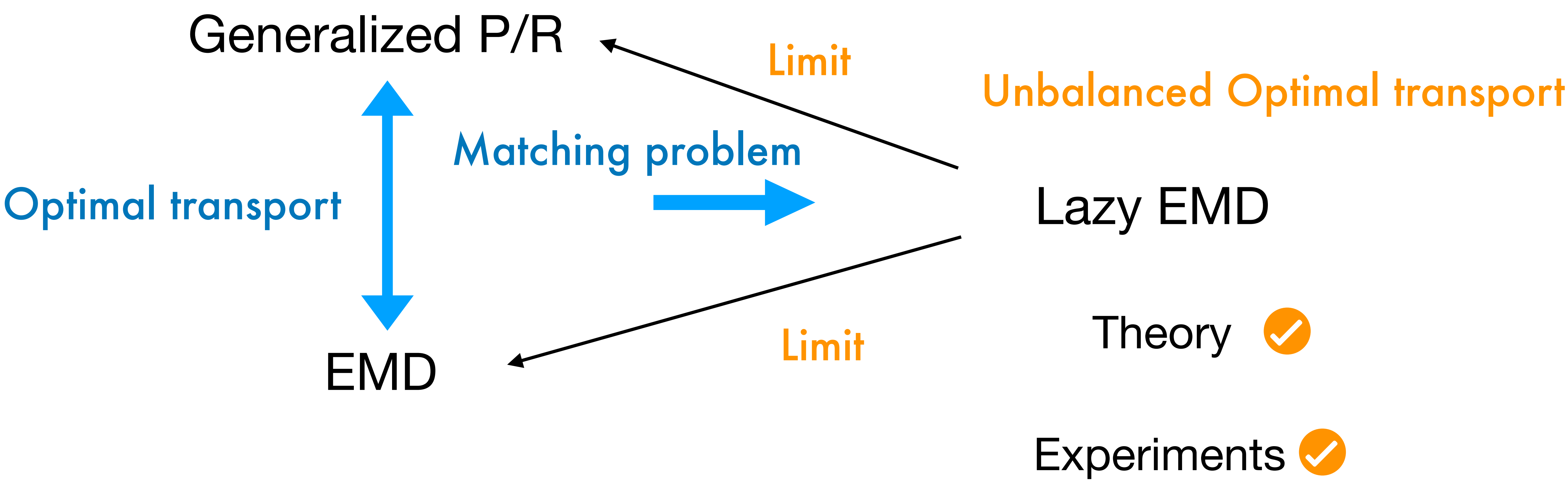
Evaluation: WMT Translation Benchmark

- WMT19: 193 translation systems, 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

Conclusion

Existing intrinsic metrics



Part 2

More Details

Part 2 - Outline

- 3 Key points

  - From optimal transport problem to token matching

  - Matching problems and evaluation

  - Why the word ‘Lazy’ ?

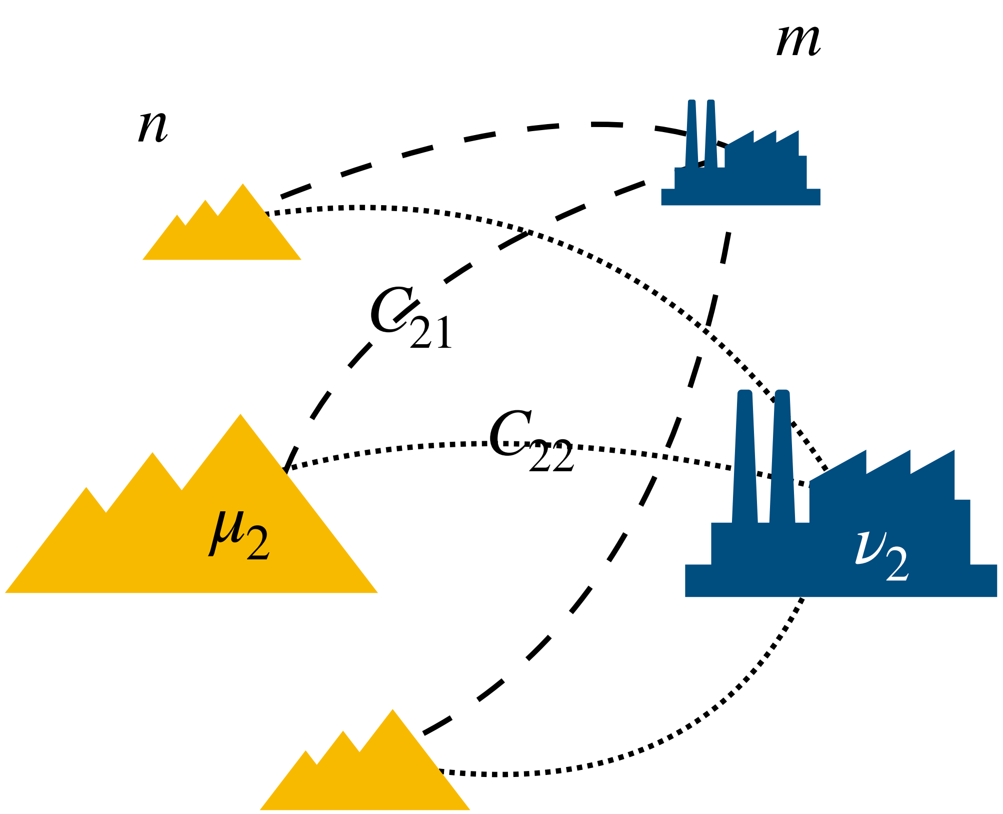
- Our Demo: visualize intrinsic metrics

  - Example

1

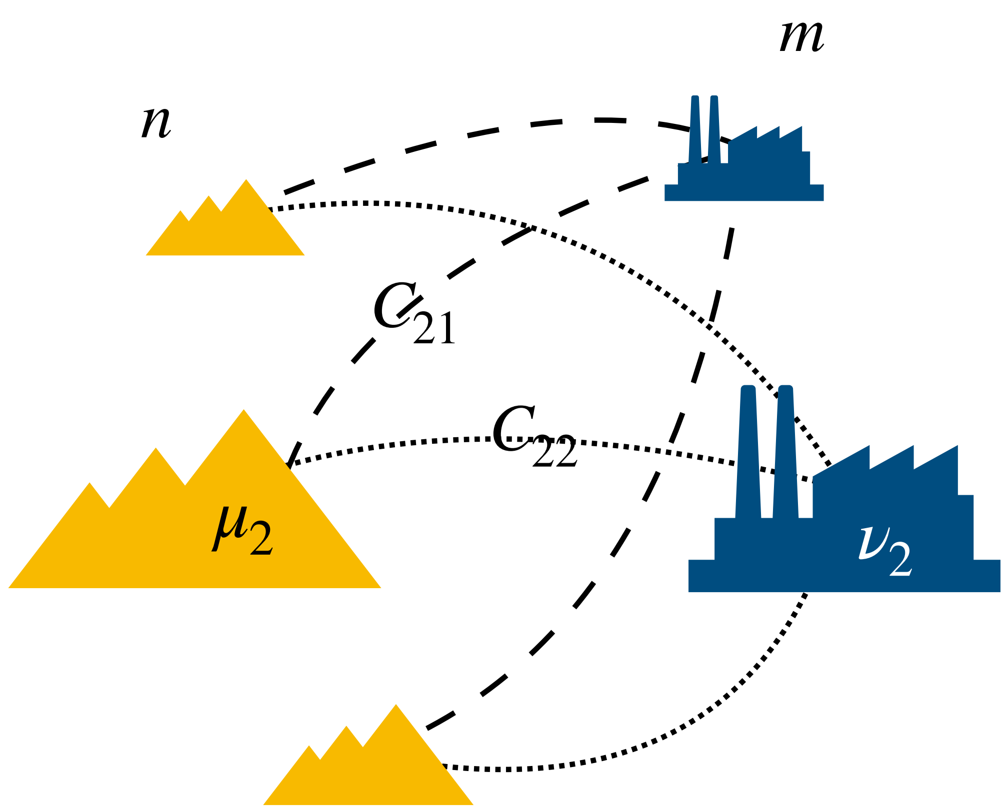
From Optimal Transport to Token Matching

Optimal Transport Problem



- Earth of mass μ_i on site i
- Requirements of mass ν_j of factory j
- Transport cost from i to j : C_{ij}
- Make the transport plan, minimize the total cost.

Optimal Transport Problem



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

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

Solution $P*$: optimal transport plan

EMD: Bilateral

token distance matrix

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

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

cand./ref. token weights

$\xrightarrow{P^*}$

$$EMD = \langle C, P^* \rangle$$

Generalized Precision/Recall: Unilateral

1 - S (similarity matrix)

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

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

cand. token weights

P_p^* →

Generalized precision

$$P = \langle S, P_p^* \rangle$$

Generalized Precision/Recall: Unilateral

1 - S (similarity matrix)

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

$$s.t. \quad \mathbf{P}^T \mathbb{1}_n = \boldsymbol{\nu}.$$

ref. token weights

$\xrightarrow{P_r^*}$

Generalized recall

$$R = \langle S, P_r^* \rangle$$

Different HARD constraints

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

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

$$s.t. \ \mathbf{P}\mathbb{1}_m = \boldsymbol{\mu} \longrightarrow P = \langle S, P_p^* \rangle$$

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

P_{ij} : Matching weight of token i, j

2

Matching Problems and Evaluation

GOOD match?

P_{ij} : how much the similarity of token pair (i, j) is considered in computing the final score.

In traditional evaluation measures like BLEU, ROUGE, the problem is the stiffness on matching
—— only words lexically similar can be matched.

However in embedding-based measures, the problem is the flexibility
—— ANY two words can be matched !

GOOD match?

P_{ij} : how much the similarity of token pair (i, j) is considered in computing the final score.

What kind of match is bad?

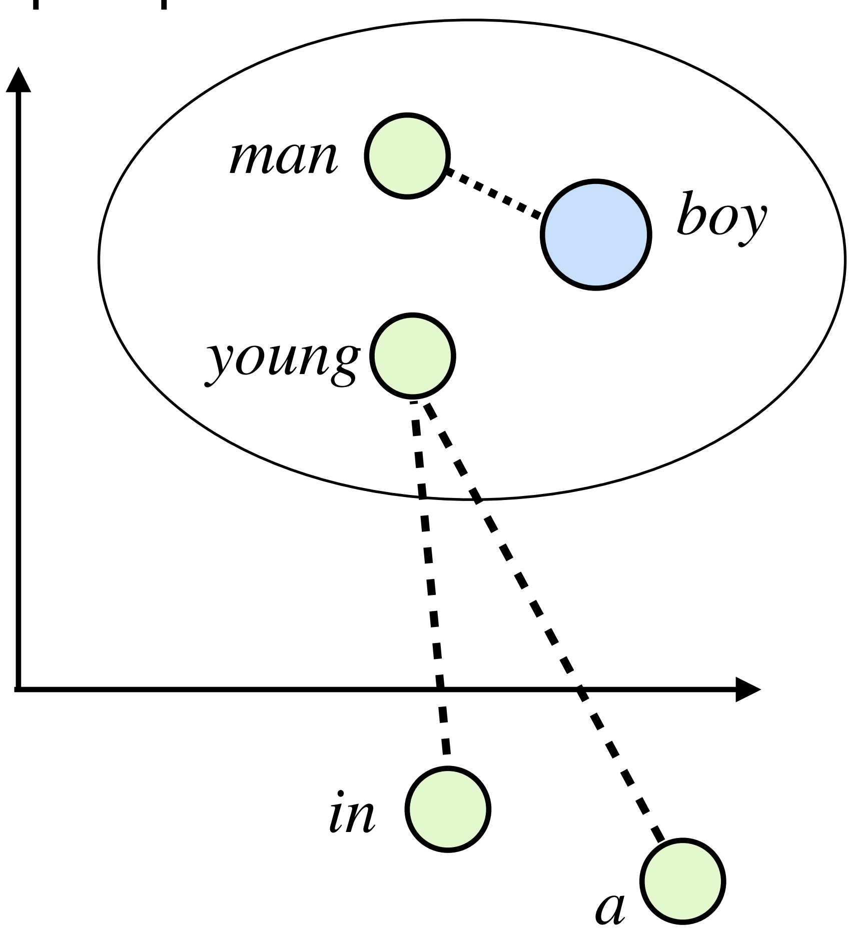
1. Incomplete matching
2. Noisy matching

HARD constraints, BAD match

Reference: The young man in a slicker.

Candidate: The boy in a coat

paraphrase



1. Incomplete matching
2. Noisy matching

Unilateral: nearest neighbor

Bilateral:
ideal only when $w_{man} + w_{young} = w_{boy}$

3

Why the word 'Lazy' ?

OT with Soft Constraints

- Unbalanced Optimal Transport Problem

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

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

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

marginal deviation,
by KL divergence

OT with Soft Constraints

- Unbalanced Optimal Transport Problem

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

control how much the corresponding
marginal deviation is penalized

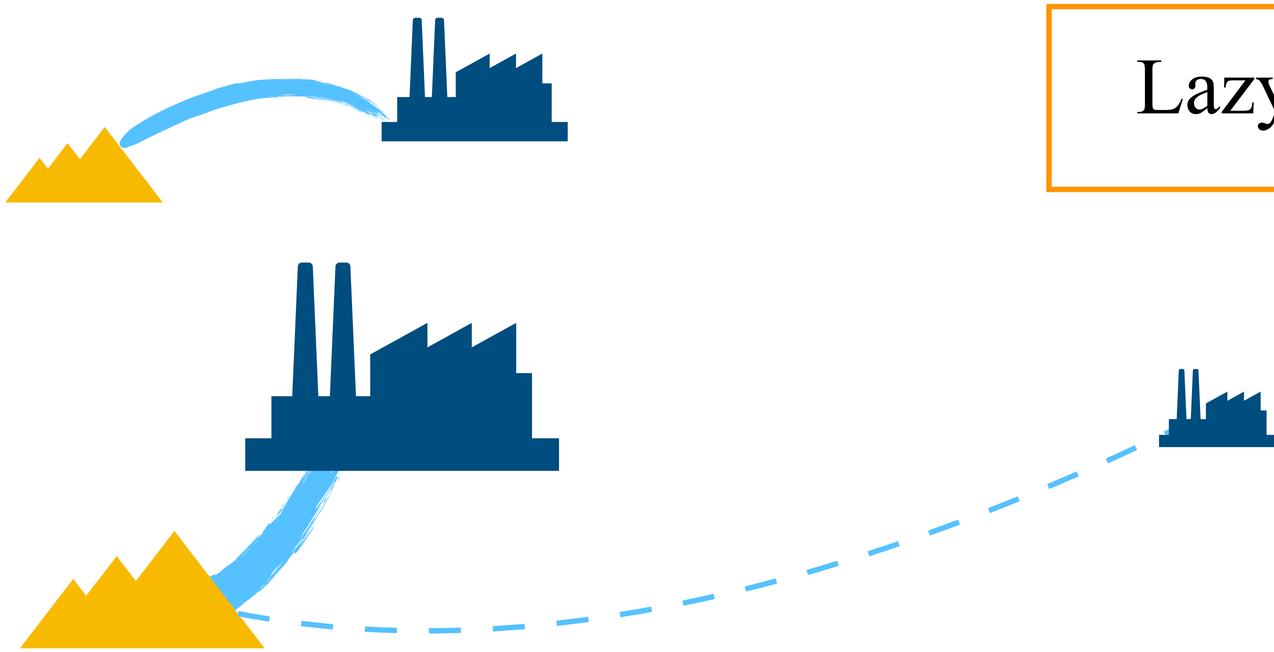
Lazy Earth Mover's Distance

- Unbalanced Optimal Transport Problem

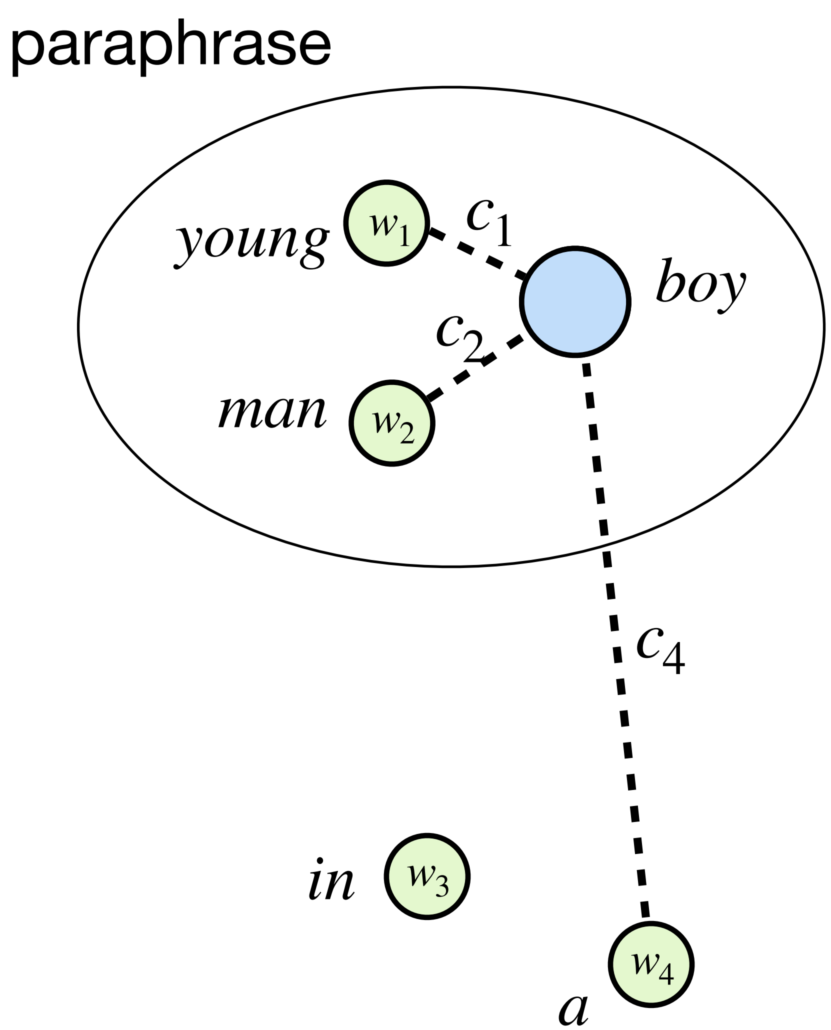
$$\min_{\mathbf{P}\in\mathbb{R}_+^{n\times m}} \langle \mathbf{C}, \mathbf{P}\rangle + \lambda_c \mathrm{KL}(\mathbf{P}\mathbb{1}_m|\boldsymbol{\mu}) + \lambda_r \mathrm{KL}(\mathbf{P}^T\mathbb{1}_n|\boldsymbol{\nu}).$$

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

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



Lazy matching $\mathbf{P}^*_{\lambda_c,\lambda_r}$



Matching weight

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

Demo:

Compare intrinsic metrics!

Demonstration: Compare Intrinsic Metrics !

- Choose the encoder
- Explore the similarity matrix
- Get evaluation scores under different metrics
- Explore their matching weights

Thanks for your attention !

Resources: https://github.com/Beastlyprime/lazy_emd

TRY OUR DEMO!