Automatique Evaluation of Natural Language Generation using Measures of Similarity

Pierre Colombo

LEYA Lab 16/12/2021

What is automatic evaluation?

Problem:

S_1: The weather is cold today.

S_2: It is freezing today

0.8

Similar

S_1: I like those cats.

S_2: It is freezing today

0.1

Dissimilar

Goal:

Building a metric m

$$m : \mathcal{S} \times \mathcal{S} \rightarrow [0,1]$$

$$(S_1, S_2) \rightarrow m(S_1, S_2)$$

Success Criterion:

When do we know that m **is good?**

Correlation with human scores

Koehn 2009; Specia, Raj, and Turchi 2010; Chatzikoumi 2020

Importance of Evaluation of NLG

Why is automatic evaluation popular?

1. Cheap: compared to human evaluation.
2. Fast: you can label “instantaneously”.
3. Reproductible: two sentences always get the same score.
4. Easy to use (e.g no annotator training, no form design).

Karpinska et al. 2021

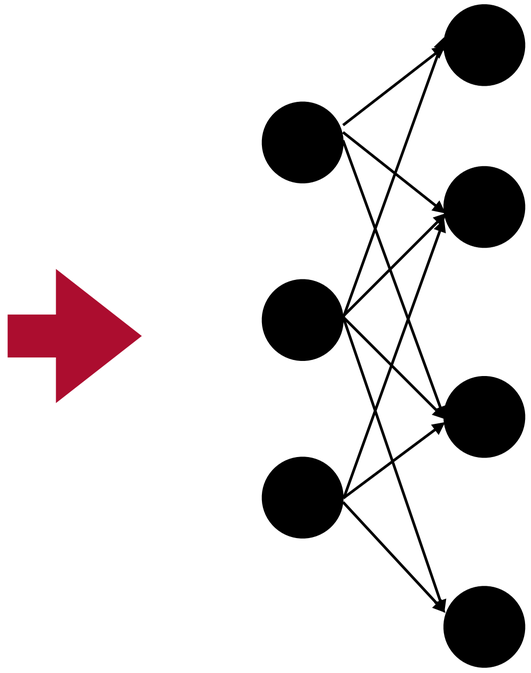
Why do we need evaluation of NLG?

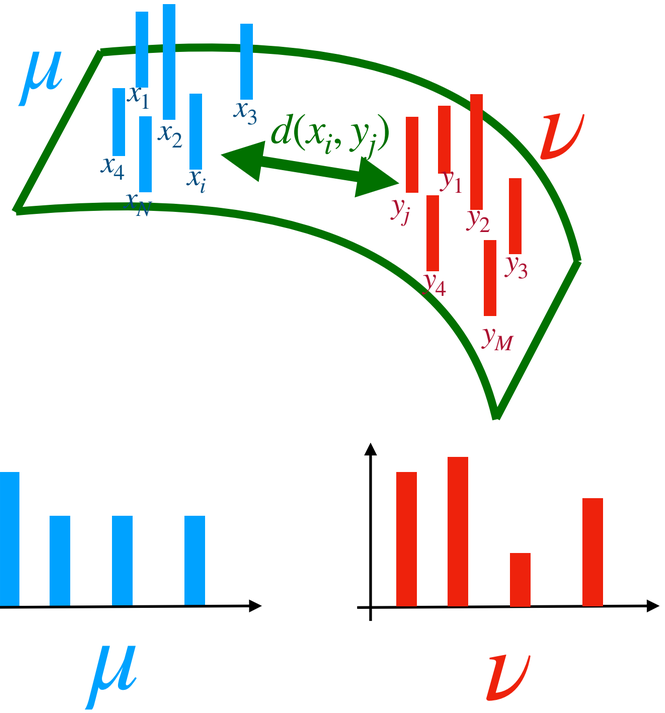
1. Debug NLG systems without annotators.
2. Improve learning of systems by deriving new losses.
3. Compare different systems.

Statistical Measures of Similarity

High dimensional data

Hello, Chicago.
If there is anyone out there who still doubts that America is a place where all things are possible, who still wonders if the dream of our founders is alive in our time, [….].
Yes we can!





Input Text | **Neural Network** | **Soft Probabilities**

When working with Neural Networks, we need to compare probability distributions

The statistical measures are a tool measure this similarity/dissimilarity between probability distribution!

Outline

1. Context: examples of problems, evaluation of automatic evaluation

1.1 Types of Automatic Evaluation of NLG

1.2 Evaluation of automatic evaluation

2. Reference Based Automatic Evaluation

2.1 DepthScore

Guillaume Staerman, Pavlo Mozharovskyi, **Pierre Colombo**, Stephan Clémençon, Florence d'Alché-Buc. A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions.

2.2 BaryScore

Pierre Colombo, Guillaume Staerman, Chloé Clavel, Pablo Piantanida. Automatic Text Evaluation through the Lens of Wasserstein Barycenters. **EMNLP 2022 (oral)**

2.2 InfoLM

Pierre Colombo, Chloé Clavel and Pablo Piantanida. InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation. **AAAI 2022**

3. Conclusions

1. Context: examples of problems, evaluation of automatic evaluation.

Reference based vs reference free evaluation

$$m : \mathcal{S} \times \mathcal{S} \rightarrow [0,1]$$

$$(S_1, S_2) \rightarrow m(S_1, S_2)$$

What are S_1 and S_2 ?

Reference based

I like running outside !

System

S_1 J'aime courir à l'extérieur !

Human

S_2 J'apprécie courir dehors !

Reference free

S_1 I like running outside !

S_2 J'aime courir à l'extérieur !

Sentence level vs word level scores

Sentence Level Scores

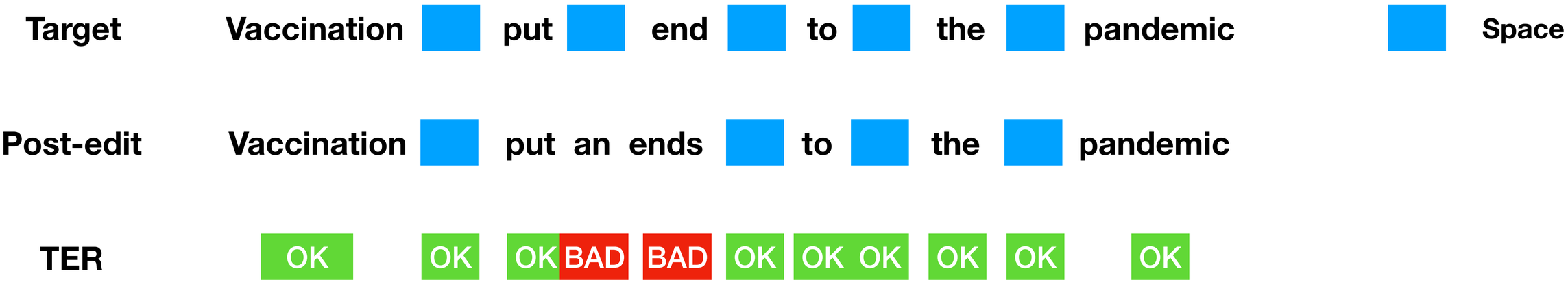
Goal: Building a metric m

$$m : \mathcal{S} \times \mathcal{S} \rightarrow [0,1]$$

$$(S_1, S_2) \rightarrow m(S_1, S_2)$$

Word Level Scores: Word-level quality estimation

Ding et al. 2021 Rei et al. 2021



Metric Evaluation

Success Criterion: **Correlation with human score**

Notations

S systems

N texts

R_i **i-th reference**

C_i^j **i-th text candidate generated by j-th system**

$h(C_i^j)$ **human score**

$$Scores = \begin{bmatrix} & M_1 & M_2 & \cdots & M_{|M|} & h_1 & h_2 & \cdots & h_{|H|} \\ S_1 \to C_1^{S_1} & & & & & & & & \\ \cdots & & & & & & & & \\ S_1 \to C_N^{S_1} & & & & & & & & \\ \cdots & & & & & & & & \\ \cdots & & & & & & & & \\ S_{|S|} \to C_1^{S_{|S|}} & & & & & & & & \\ \cdots & & & & & & & & \\ S_{|S|} \to C_N^{S_{|S|}} & & & & & & & & \end{bmatrix}$$

$M_{i_0}(C_j^i, R_j)$ $h_{k_0}(C_j^i, R_j)$

Can the metric be used to compare the performance of two systems?

Can the metric be used as a loss or reward of a system?

Metric Evaluation: system level aggregation

Can the metric be used to compare the performance of two systems?

$$K_{sys} = K(M^{sy}, H^{sy})$$

$$M^{sy} = \left[\frac{1}{N} \sum_{i=1}^{N} m(R_i, C_i^1), \cdots, \frac{1}{N} \sum_{i=1}^{n} m(R_i, C_i^s) \right]$$

$$H^{sy} = \left[\frac{1}{N} \sum_{i=1}^{N} h(C_i^1), \cdots, \frac{1}{N} \sum_{i=1}^{N} h(C_i^S) \right]$$

$$Scores = \begin{bmatrix} & M_1 & M_2 & \cdots & M_{|M|} & h_1 & h_2 & \cdots & h_{|H|} \\ S_1 \rightarrow C_1^{S_1} & & & & & & & & \\ \cdots & & & & & & & & \\ S_1 \rightarrow C_N^{S_1} & \Sigma & & & & \Sigma & & & \\ \cdots & & & & & & & & \\ S_i \rightarrow C_1^{S_i} & & & & & & & & \\ \cdots & & & & & & & & \\ S_i \rightarrow C_N^{S_i} & \Sigma & & & & \Sigma & & & \\ \cdots & & & & & & & & \\ \cdots & & & & & & & & \\ S_{|S|} \rightarrow C_1^{S_{|S|}} & & & & & & & & \\ \cdots & & & & & & & & \\ S_{|S|} \rightarrow C_N^{S_{|S|}} & \Sigma & & & & \Sigma & & & \end{bmatrix}$$

System Aggregation !
Compare vector of length S

Metric Evaluation: text level Aggregation

Can the metric be used as a loss or reward of a system?

$$K_{text} = \frac{1}{N}\sum_{i=1}^{N} K(M_i^{text}, H_i^{text})$$

$$H_i^{text} = \left[h(C_i^1), \cdots, h(C_i^S)\right]$$

$$M_i^{text} = \left[m(R_i, C_i^1), \cdots, m(R_i, C_i^S)\right]$$

$$Scores = \begin{bmatrix} & M_1 & M_2 & \cdots & M_{|M|} & h_1 & h_2 & \cdots & h_{|H|} \\ S_1 \to C_1^{S_1} & \blacksquare & & & & \blacksquare & & & \\ \cdots & & & & & & & & \\ S_1 \to C_N^{S_1} & & & & & & & & \\ \cdots & & & & & & & & \\ S_i \to C_1^{S_i} & \blacksquare & & & & \blacksquare & & & \\ S_i \to C_N^{S_i} & & & & & & & & \\ \cdots & & & & & & & & \\ S_{|S|} \to C_1^{S_{|S|}} & \blacksquare & & & & \blacksquare & & & \\ \cdots & & & & & & & & \\ S_{|S|} \to C_N^{S_{|S|}} & & & & & & & & \end{bmatrix}$$

Text Aggregation !
Averaged correlation

A summary so far …..

What are the different settings?

Reference Based/ Reference Free

Different Granularities (word vs sentence)

How to evaluate a metric?

Text Level Correlation

Can the metric be used as a loss or reward of a system?

System Level Correlation

Can the metric be used to compare the performance of two systems?

Let’s speak about the different metrics for reference-based evaluation!

2. Reference Based Automatic Evaluation of NLG

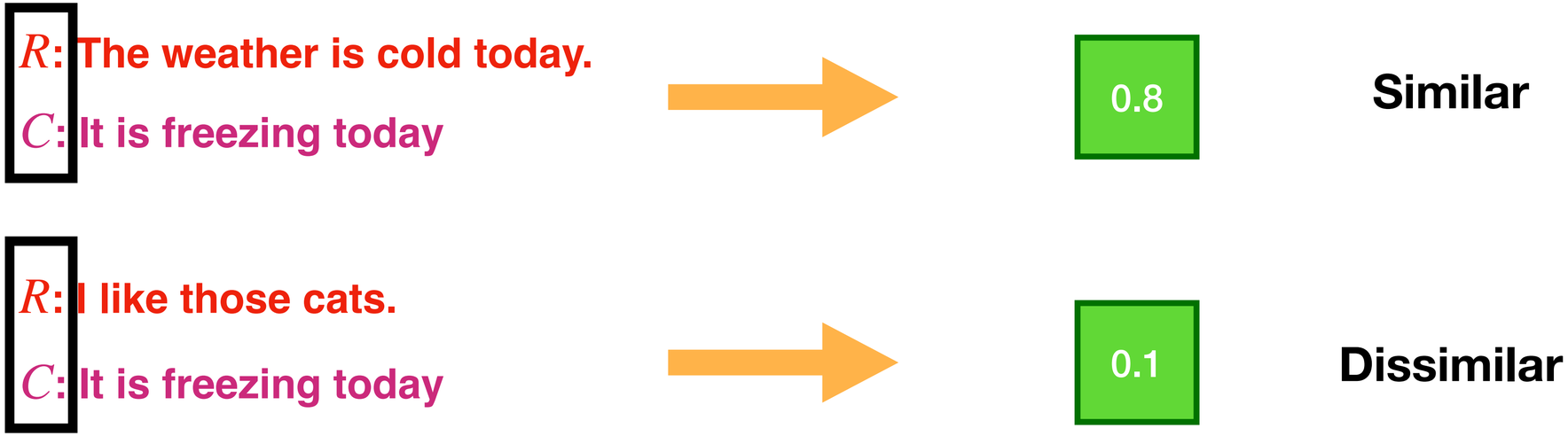
Pierre Colombo, Guillaume Staerman, Chloé Clavel, Pablo Piantanida. Automatic Text Evaluation through the Lens of Wasserstein Barycenters. (oral) EMNLP 2021

Pierre Colombo, Chloé Clavel and Pablo Piantanida. InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation. AAAI 2022

Guillaume Staerman, Pavlo Mozharovskyi, Pierre Colombo, Stephan Clémençon, Florence d'Alché-Buc. A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions. Submitted

Existing Metrics for Reference Based NLG

Goal



Edit Based **N-gram Based** **Embedding Based**

Existing Methods

Edit Based

Snover et al. 2006

Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)
sailor -> sailir (S)
sailir -> sailin (S)
sailin -> sailing (I)

Distance is 4 !

N-gram Based

Papineni et al. 2002

C : I like these very nice pies !

R : I like those cakes !

Unigrams

C : I like these very nice pies !

R : I like those cakes !

Bigrams

C : I like these very nice pies !

R : I like those cakes !

Embedding Based

Word Mover distance

Kusner et al. 2015

BertScore

Zhang et al. 2019

MoverScore

Zhao et al. 2019

Sentence Mover

Clark et al. 2019

Existing Methods

Edit Based

Snover et al. 2006

Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)
sailor -> sailir (S)
sailir -> sailin (S)
sailin -> sailing (I)

Distance is 4 !

N-gram Based

Papineni et al. 2002

C : I like these very nice pies !

R : I like those cakes !

Unigrams

C : I like these very nice pies !

R : I like those cakes !

Bigrams

C : I like these very nice pies !

R : I like those cakes !

BertScore & DepthScore

Embedding Based

Word Mover distance

Kusner et al. 2015

BertScore

Zhang et al. 2019

MoverScore

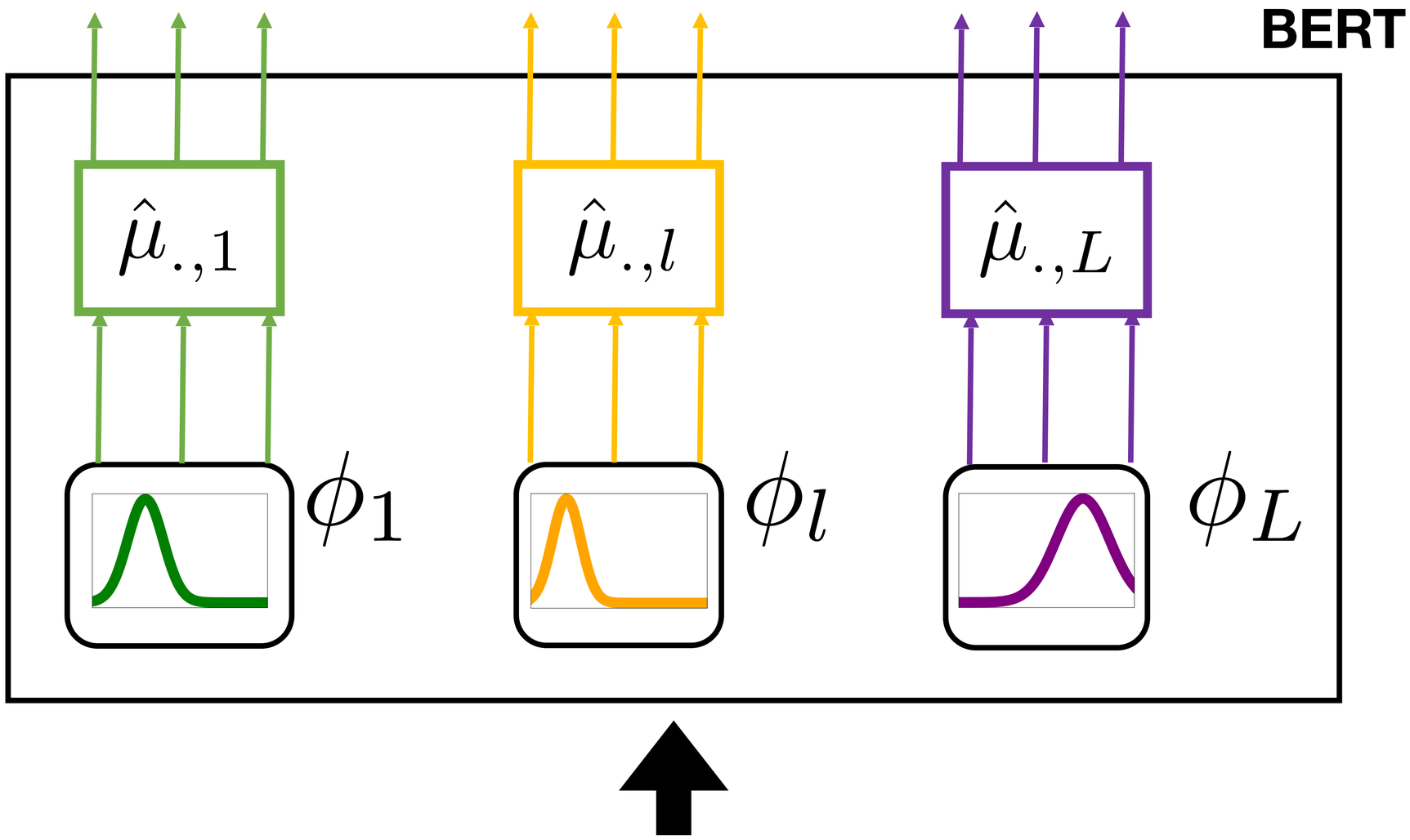
Zhao et al. 2019

Sentence Mover

Clark et al. 2019

A recall on the hidden space of BERT

Delvin et al. 2019

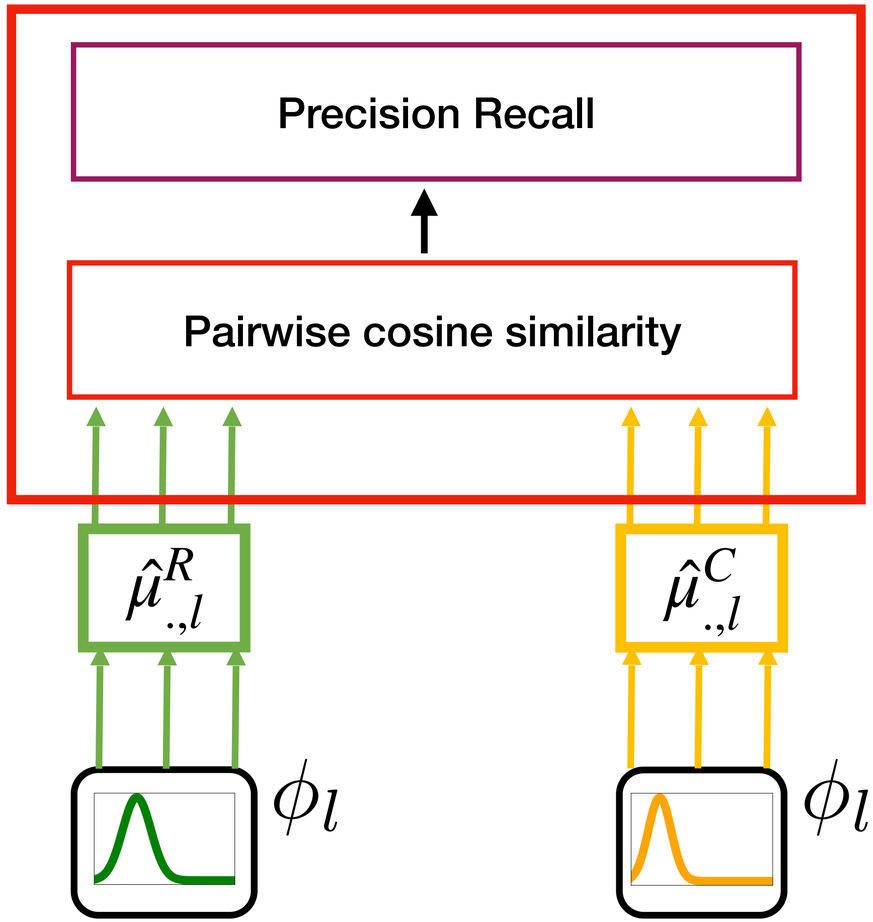


1. “Contextual Embedding”
2. 12 Layers (L=12)
3. H = 768
4. Max length = 512

The weather is cold today ! I am dying :’(!

BertScore

Zhang et al. 2019



R: The weather is cold today

C: It is freezing this morning

This is a distance between two empirical distributions !

Advantage

1. Deal with paraphrases
2. Include "semantic"

Limitations

1. Use only one layer
2. Use arbitrary sequence of operation

DepthScore

Guillaume Staerman, Pavlo Mozharovskyi, Pierre Colombo, Stephan Clémençon, Florence d'Alché-Buc.
A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions. Submitted

A glance on data-depths

G. Staerman, P. Mozharovskyi, P. Colombo, S. Clémençon, F. d'Alché-Buc. A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions.

Let's build an alternative distance between distribution !

Definition

Data depth are non parametric statistic which measures the centrality of any element of a $\mathcal{X}$ w.r.t. a probability distribution $\mathcal{P}_{\mathcal{X}}$.

$$D : \mathcal{X} \times \mathcal{P}(\mathcal{X}) \to [0,1]$$
$$\left(x, \rho\right) \to D(x, \rho) = D_{\rho}(x)$$

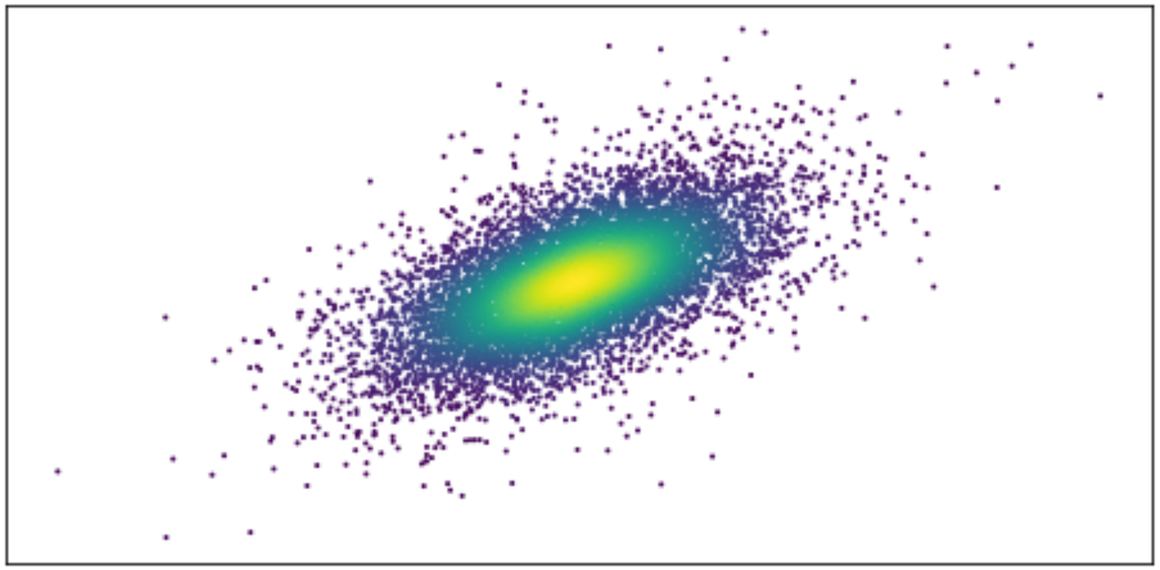
Properties

Affine Invariant

Monotonous on rays

Maximal at symmetry center

Vanishing at infinity

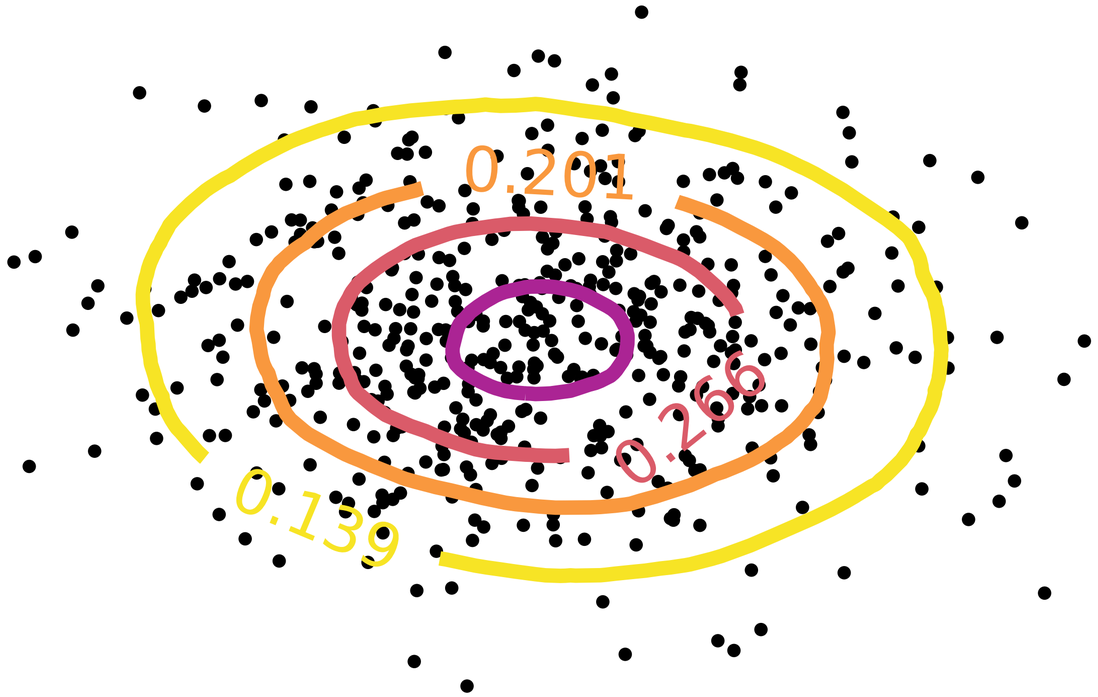


Staerman et al. 2020

From data-depths to a distance

Definition **For any $\alpha \in [0,1]$, the associated α-depth region of a depth function is defined as:**

$$D_\rho^\alpha = \left\{ x \in \mathbb{R}^d, D_\rho(x) \geq \alpha \right\}$$



Notes

Depth regions are nested and they generalise the notion of quantiles to a multivariate distribution.

A pseudo metric based on depth-trimmed regions

Definition $DR_{p,\varepsilon}$ **is a discrepancy measure between** μ, ν

$$DR_{p,\varepsilon}(\mu,\nu) = \left(\int_{\varepsilon}^{1} d_{\mathscr{H}} \left(D_{\mu}^{\alpha}, D_{\nu}^{\alpha} \right)^{p} d\alpha \right)^{\frac{1}{p}}$$

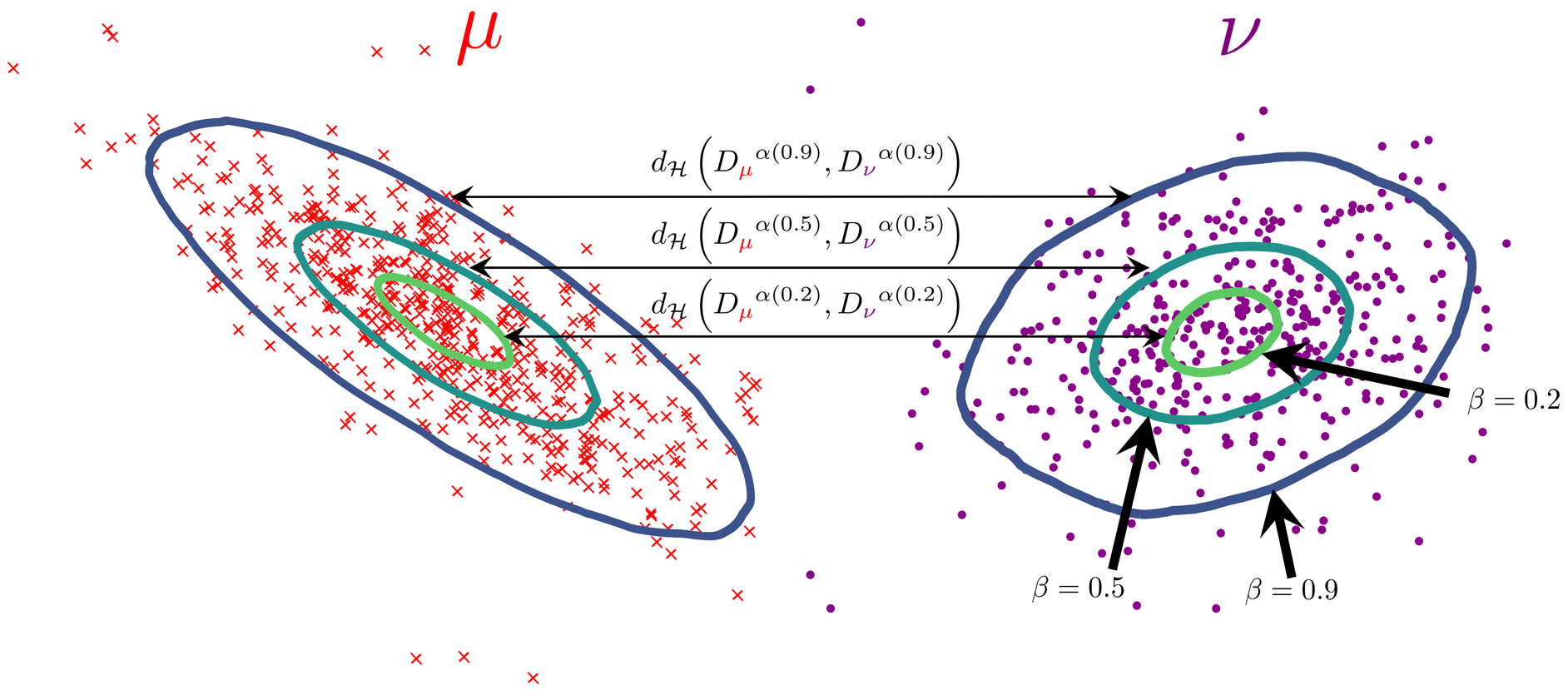
$\epsilon \in (0,1]$ $\quad p \in (0,\infty)$ $\quad d_{\mathscr{H}}$

Properties

Robustness

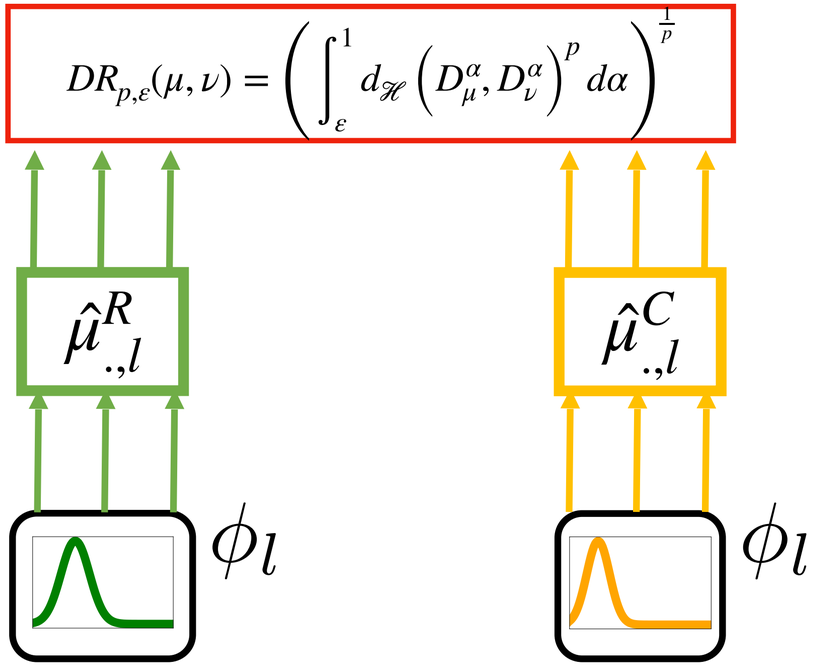
Isometry invariance (translations, rotations)

Positive, symmetric, triangular inequality



DepthScore

G. Staerman, P. Mozharovskyi, P. Colombo, S. Clémençon, F. d'Alché-Buc. A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions.



R: The weather is cold today

C: It is freezing this morning

Advantage

1. Deal with paraphrases
2. Include "semantic"

Limitations

1. Use only one layer

Results

Experimental Setting

Data2text Generation

- Results on WebNLG 2020
- Correctness / Data Coverage / Relevance
- Results on English only

Summary Generation

- Results on SummEval
- Correlation with pyramid score
- Results on English only

Summary Generation

	Abstractive			Extractive		
	r	τ	ρ	r	τ	ρ
$DR_{p,\varepsilon}$	72.1	72.1	70.1	91.5	**91.5**	69.2
Wasserstein	71.0	70.4	71.1	74.2	74.2	40.0
Sliced-Wasserstein	70.1	68.7	71.0	72.4	73.9	69.2
MMD	68.2	67.5	67.9	75.6	75.6	56.1
BertScore	71.7	71.9	72.0	70.9	72.9	**73.8**
MoverScore	72.4	71.9	73.0	76.1	76.1	47.4
ROUGE-1	**73.5**	73.0	**74.4**	72.2	74.0	69.1
ROUGE-2	73.0	**73.5**	73.0	55.1	53.2	69.1
JS-2	68.9	6.8	69.8	**92.9**	5.5	19.0

Towards Multi-layers Metrics

DepthScore & BertScore rely on a single layer !

Why are they design this way?

Before single layer embedding	Glove, Word2vec
Many prior work on single layer embeddings	WMD

What granularity is used ?

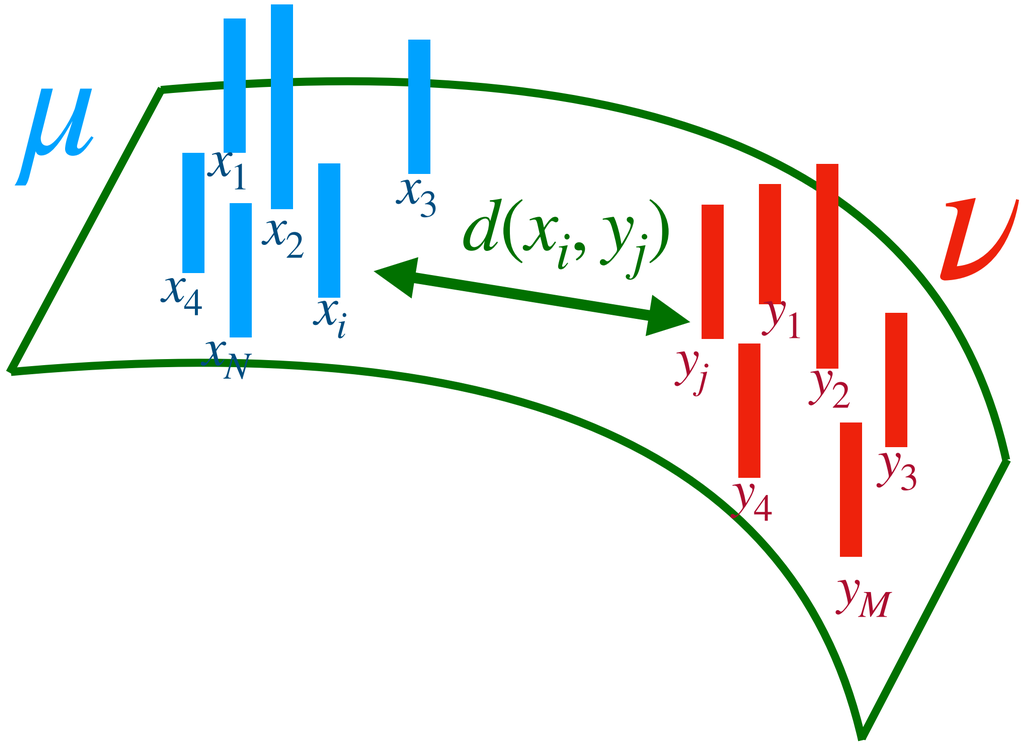
Word Level distance vs Sentence Level Distance	SentenceMover

BaryScore

Pierre Colombo, Guillaume Staerman, Chloé Clavel, Pablo Piantanida. Automatic Text Evaluation through the Lens of Wasserstein Barycenters. (oral) EMNLP 2021

Optimal Transport

Goal Compute distance between probability measures (μ,ν)



Input Discret Measures

$$\nu = \sum_{j=1}^{M} \beta_j \delta_{x_j} \qquad \mu = \sum_{i=1}^{N} \alpha_i \delta_{x_i}$$

Cost Matrix

$$C = \begin{pmatrix} d(x_1, y_1) & \cdots & d(x_1, y_M) \\ \cdots & \cdots & \cdots \\ d(x_N, y_1) & \cdots & d(x_N, y_M) \end{pmatrix}$$

Transport Plan

$$\Pi = \begin{pmatrix} \pi_{11} & \cdots & \pi_{1M} \\ \cdots & \cdots & \cdots \\ \pi_{N1} & \cdots & \pi_{NM} \end{pmatrix} \begin{matrix} \rightarrow \beta_1 \\ \\ \rightarrow \beta_N \end{matrix}$$

α_1 α_M

Wasserstein Distance

$$OT(\nu, \mu) = \min_{\Pi} \sum_{ij} C_{i,j} \times \Pi_{i,j}$$

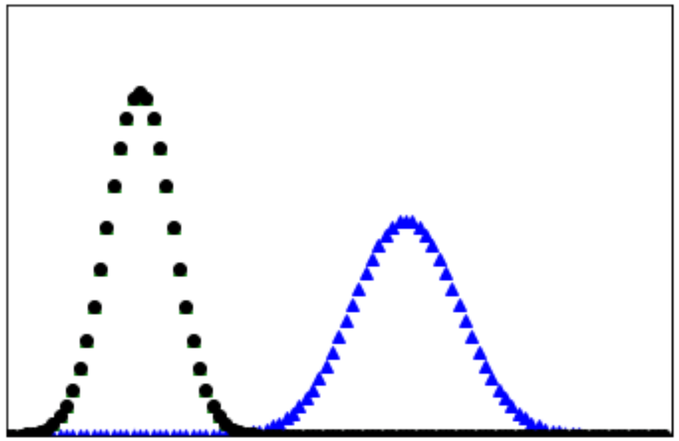
$$\Pi 1 = \alpha, \Pi^T 1 = \beta$$

Wasserstein Barycenters

Euclidean Interpolation

$$\nu = \sum_{i=1}^{N} \alpha_i l_2(\mu_i, \mu)$$

$$\nu = \alpha_i l_2(\mu_i, \mu) + (1 - \alpha_i) l_2(\mu_i, \mu)$$



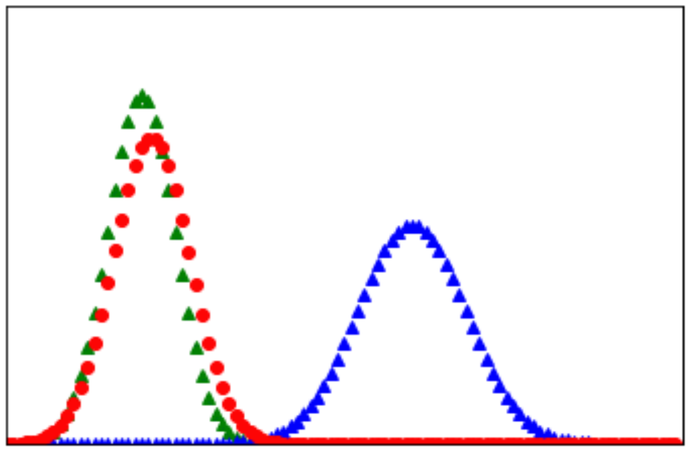
Do not look like a gaussian !

α_i **Varies**

Wasserstein Interpolation

$$\nu = argmin_{\mu} \sum_{i=1}^{N} \alpha_i W(\mu_i, \mu)$$

$$\nu = argmin_{\mu} \alpha_i W(\mu_i, \mu) + (1 - \alpha_i) W(\mu_i, \mu)$$



Preserve the gaussian!

BaryScore

Reference: R **Candidate:** C

Goal : metric $m : (C, R) \mapsto m(C, R) \in \mathbb{R}_+$

Algorithm

1. Find the Wasserstein barycentric distributions of BERT layers for C and R

2. Evaluate these barycentric distributions using the Wasserstein distance.

Algorithm 1 `BaryScore`

INPUT: $C = \{\omega_1^c, \ldots, \omega_{n_c}^c\}$, $R = \{\omega_1^r, \ldots, \omega_{n_r}^r\}$, $(\phi_1, \ldots, \phi_L)$ pre-trained layers from BERT or ELMo.

Compute layers embeddings:
$\phi_\ell(C)$ and $\phi_\ell(R)$ for every $1 \leq \ell \leq L$.

Compute measures: $\{\hat{\mu}_{C,\ell}, \hat{\mu}_{R,\ell}\}_{\ell=1}^L$.

Compute Wasserstein barycenters:

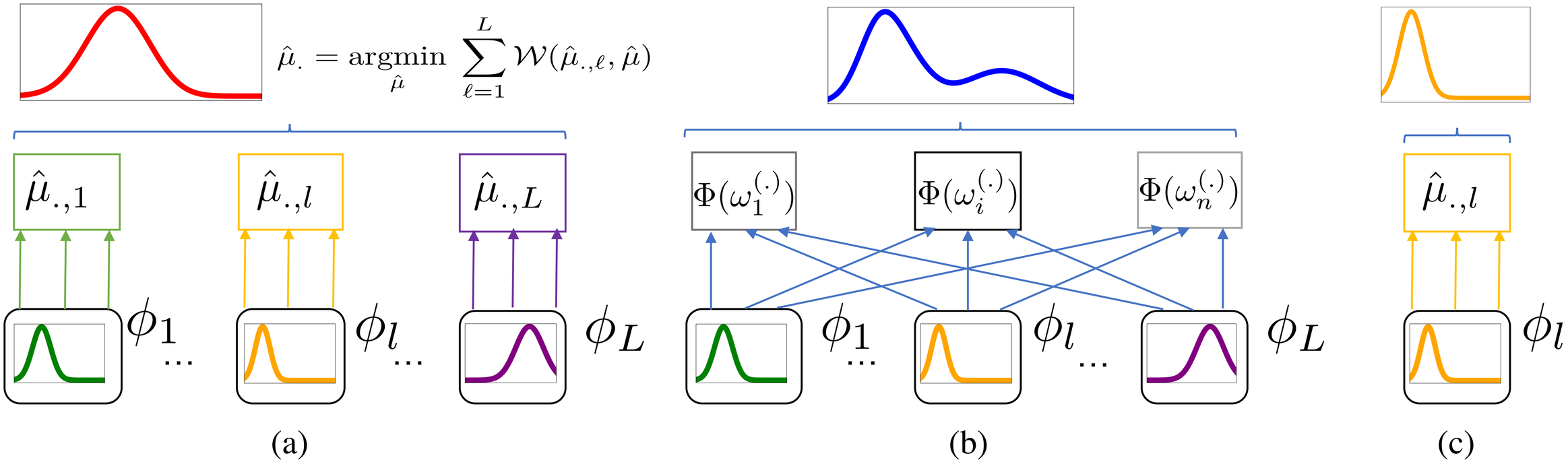
$$\hat{\mu}_C = \operatorname*{argmin}_{\hat{\mu}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{C,\ell}, \hat{\mu}),$$

$$\hat{\mu}_R = \operatorname*{argmin}_{\hat{\mu}} \sum_{\ell=1}^{L} \mathcal{W}(\hat{\mu}_{R,\ell}.\hat{\mu}),$$

OUTPUT: $\mathcal{W}(\hat{\mu}_R, \hat{\mu}_C)$.

BaryScore vs BertScore vs MoverScore

Comparison between aggregation functions



BaryScore **MoverScore** **BertScore**

Experimental Setting

Machine Translation

- Results on WMT17/WMT18
- All metrics are measures on en only
- Pairs includes cs-en de-en ru-en fi-en ro-en tr-en

Data2text Generation

- Results on WebNLG 2020
- Correctness / Data Coverage / Relevance
- Results on English only

Image Captioning

- Results on MSCOCO
- Results on English only

Summary Generation

- Results on SummEval
- Correlation with pyramid score
- Results on English only

Results

Task

(John_Blaha birthDate 1942_08_26)
(John_Blaha birthPlace San_Antonio)
(John_E_Blaha job Pilot)

John Blaha, born in San Antonio on 1942-08-26, worked as a pilot

Criterion

- Correctness
- Data coverage
- Relevance

	Correctness			Data Coverage			Relevance		
Metric	r	ρ	τ	r	ρ	τ	r	ρ	τ
Correct	100.0	100.0	100.0	97.6	85.2	73.3	99.1	89.7	75.0
DataC	85.2	97.6	73.3	100.0	100.0	100.0	96.0	93.8	81.6
Relev	89.7	99.1	75.0	96.0	93.8	81.6	100.0	100.0	100.0
BaryS	**91.7**	**90.0**	**78.3**	**87.8**	78.2	61.6	**89.4**	**82.6**	70.0
BaryS$^+$	90.5	89.5	76.6	87.7	**85.0**	**70.0**	89.2	86.4	71.6
BertS	85.5	83.4	73.3	74.7	68.2	53.3	83.3	79.4	65.0
MoverS	84.1	84.1	73.3	78.7	66.2	53.3	82.1	77.4	65.0
BLEU	77.6	66.3	60.0	55.7	50.2	36.6	63.0	65.2	51.6
R-1	80.6	65.0	65.0	76.5	76.3	60.3	64.3	69.2	56.7
R-2	73.6	63.3	58.3	54.7	43.1	35.0	62.0	60.8	46.7
R-WE	60.9	73.4	60.0	40.2	58.2	40.1	49.9	64.1	48.3
METEOR	86.5	66.3	70.0	77.3	50.2	46.6	82.1	65.2	58.6
TER	79.6	78.3	58.0	69.7	58.2	38.0	75.0	70.2	**77.6**

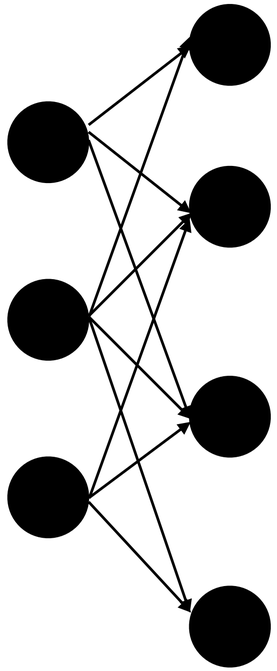
Correlation score for different coefficient Pearson r, Spearman ρ and Kendall τ

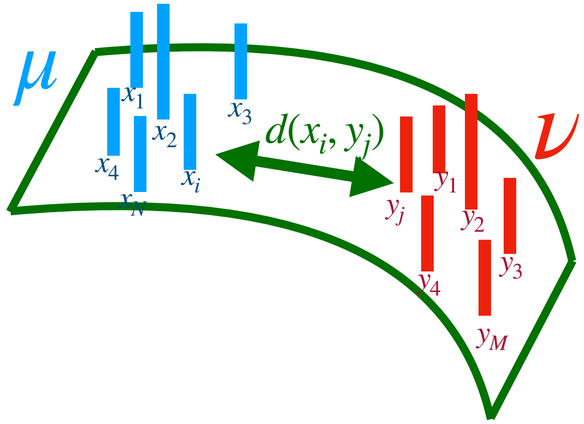
InfoLM

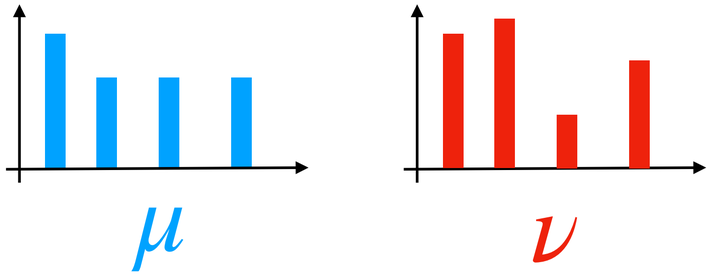
Pierre Colombo, Chloé Clavel and Pablo Piantanida. InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation. AAAI 2022

Statistical Measures of Similarity

Hello, Chicago.
If there is anyone out there who still doubts that America is a place where all things are possible, who still wonders if the dream of our founders is alive in our time, [….].
Yes we can!







Input Text

Neural Network

High dimensional data

Soft Probabilities

Existing Methods

Edit Based

Snover et al. 2006

Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)
sailor -> sailir (S)
sailir -> sailin (S)
sailin -> sailing (I)

Distance is 4 !

N-gram Based

Papineni et al. 2002

C : I like these very nice pies !

R : I like those cakes !

Unigrams

C : I like these very nice pies !

R : I like those cakes !

Bigrams

C : I like these very nice pies !

R : I like those cakes !

InfoLM

Embedding Based

Word Mover distance

Kusner et al. 2015

BertScore

Zhang et al. 2019

MoverScore

Zhao et al. 2019

Sentence Mover

Clark et al. 2019

Assumptions for InfoLM

Goal Compute a similarity score between R and C.

Tools Use a pretrained MLM

MLM predicts a distribution over Ω

$$p_\Omega(\cdot \mid [R]^i)$$

Use a measure of information

$$\mathcal{I} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$$

Name	Notation	Domain	Expression
α-divergence (Csiszár 1967)	$\mathcal{D}_\alpha$	$\alpha \notin \{0,1\}$	$\frac{1}{\alpha(\alpha-1)}(1 - \sum q_i^{1-\alpha} p_i^\alpha)$
γ divergence (Fujisawa and Eguchi 2008)	$\mathcal{D}_\gamma^\beta$	$\beta \notin \{0,-1\}$	$\frac{1}{\beta(\beta+1)} \log \sum p_i^{\beta+1} + \frac{1}{\beta+1} \log \sum q_i^{\beta+1} - \frac{1}{\beta} \log \sum p_i q_i^\beta$
AB Divergence (Cichocki, Cruces, and Amari 2011)	$\mathcal{D}_{sAB}^{\alpha,\beta}$	$(\alpha,\beta) \in (\mathbb{R}^*)^2$ $\beta + \alpha \neq 0$	$\frac{1}{\beta(\beta+\alpha)} \log \sum p_i^{\beta+\alpha} + \frac{1}{\beta+\alpha} \log \sum q_i^{\beta+\alpha} - \frac{1}{\beta} \log \sum p_i^\alpha q_i^\beta$
$\mathcal{L}_1$ distance	$\mathcal{L}_1$		$\sum \lvert p_i - q_i \rvert$
$\mathcal{L}_2$ distance	$\mathcal{L}_2$		$\sqrt{\sum (p_i - q_i)^2}$
$\mathcal{L}_\infty$ distance	$\mathcal{L}_\infty$		$\max_i \lvert p_i - q_i \rvert$
Fisher-Rao distance	R		$\frac{2}{\pi} \arccos \sum \sqrt{p_i \times q_i}$

Intuition of InfoLM

Goal **Compute a similarity score between R and C.**

Equivalence for masked contexts $\mathcal{I} : [0,1]^{|\Omega|} \times [0,1]^{|\Omega|}$

MLM predicts a distribution over Ω

$p_\Omega(\cdot \,|\, [R]^i)$

Similar context

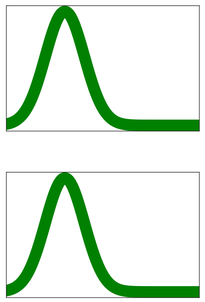
R: It is [MASK] today.

C: It is [MASK] this morning !

MLM

$p_\Omega(\cdot \,|\, [R]^2)$

$p_\Omega(\cdot \,|\, [C]^2)$



$\mathcal{I}\left(p_\Omega(\cdot \,|\, [R]^2), p_\Omega(\cdot \,|\, [C]^2)\right) \sim 0$

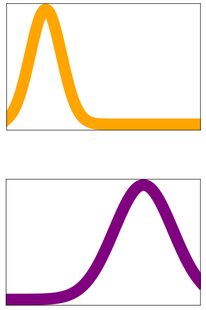
Dissimilar context

R: It is cold [MASK]

C: It is [MASK] this morning !

$p_\Omega(\cdot \,|\, [R]^3)$

$p_\Omega(\cdot \,|\, [C]^2)$



$\mathcal{I}\left(p_\Omega(\cdot \,|\, [R]^3), p_\Omega(\cdot \,|\, [C]^2)\right) \gg 0$

Context Aggregation

Goal Compute a similarity score between R and C.

How to aggregate contexts? → **Weighted Sum!**

Reference

[MASK] is cold today.

...

It is [MASK] today.

...

It is cold today [MASK]

$$P \triangleq \frac{1}{5}\sum_{k=0}^{4} \gamma_k \times p_\Omega(\cdot \mid [R]^k)$$

$$InfoLM\,(R, C) \triangleq \mathcal{I}\left(P, Q\right)$$

Candidate

[MASK] is freezing this morning !

...

It is [MASK] this morning !

...

It is freezing this morning [MASK]

$$Q \triangleq \frac{1}{6}\sum_{k=0}^{5} \gamma_k \times p_\Omega(\cdot \mid [C]^k)$$

Experimental Setting

Data2text Generation

- Results on WebNLG 2020 — Gardent et al. 2017
- Correctness / Data Coverage / Relevance — Ferreira et al. (2020)
  Fluency / Text Structure — Perez-Beltrachini et al 2016
- Results on English only

Summary Generation

- Results on SummEval — Nallapati et al. 2016); Bhandari et al. (2020)
- Correlation with pyramid score — Nenkova and Passonneau 2004
- Results on English only

Results

Task

(John_Blaha birthDate 1942_08_26)
(John_Blaha birthPlace San_Antonio)
(John_E_Blaha job Pilot)

John Blaha, born in San Antonio on 1942-08-26, worked as a pilot

Parameter Free

Metric	Correctness			Data Coverage			Fluency			Relevance			Text Structure		
	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
Correct	100.0	100.0	100.0	97.6	85.2	73.3	80.0	81.1	61.6	99.1	89.7	75.0	80.1	80.8	60.0
DataC	85.2	97.6	73.3	100.0	100.0	100.0	71.8	51.7	38.3	96.0	93.8	81.6	71.6	51.4	36.6
Fluency	81.1	80.0	61.6	71.8	51.7	38.3	100.0	100.0	100.0	77.0	61.4	46.6	99.5	99.7	98.3
Relev	89.7	99.1	75.0	96.0	93.8	81.6	77.0	61.4	46.6	100.0	100.0	100.0	77.2	61.1	45.0
TextS	80.8	80.1	60.0	71.6	51.4	36.6	99.5	99.7	98.3	77.2	61.1	45.0	100.0	100.0	100.0
$\mathcal{D}_{AB}$	88.8	**89.3**	**76.6**	**81.8**	**82.6**	**70.0**	86.6	92.0	76.6	**89.8**	**87.9**	73.3	86.6	91.4	75.0
$\mathcal{D}_{\alpha}$	88.8	**89.3**	**76.6**	**81.8**	**82.6**	**70.0**	86.6	92.0	76.6	**89.8**	**87.9**	73.3	86.6	91.4	75.0
$\mathcal{D}_{\beta}$	81.4	50.0	71.6	48.4	79.7	65.0	44.8	84.7	76.6	49.3	72.3	60.0	48.0	83.8	75.0
$\mathcal{L}_1$	75.2	33.8	61.6	32.4	53.8	40.0	22.7	83.5	73.3	32.2	57.9	45.0	25.6	83.2	71.6
$\mathcal{R}$	**89.7**	86.0	75.0	78.7	70.5	51.6	**93.3**	**95.7**	**85.3**	87.6	84.4	70.0	**92.4**	93.8	81.6
JS	79.4	81.1	70.0	69.3	75.5	60.0	89.4	91.4	75.0	81.7	70.5	60.0	91.9	91.1	73.3
BertS	85.5	83.4	73.3	74.7	68.2	53.3	92.3	95.5	85.0	83.3	79.4	65.0	91.9	**95.0**	**83.3**
MoverS	84.1	84.1	73.3	78.7	66.2	53.3	91.2	92.1	78.3	82.1	77.4	65.0	90.1	91.4	76.3
BLEU	77.6	66.3	60.0	55.7	50.2	36.6	89.4	90.5	78.3	63.0	65.2	51.6	88.5	89.1	76.6
R-1	80.6	65.0	65.0	61.1	59.6	48.3	76.5	76.3	60.3	64.3	69.2	56.7	75.9	77.5	58.3
METEOR	86.5	66.3	70.0	77.3	50.2	46.6	86.7	90.5	78.3	82.1	65.2	58.6	86.2	89.1	76.6
TER	79.6	78.3	58.0	69.7	58.2	38.0	89.1	93.5	80.0	75.0	70.2	**77.6**	89.5	91.1	78.6

Summary

Summary In this presentation we have explored the problem of NLG evaluation with a specific focus on reference based evaluation

New Metrics

- DepthScore
- BaryScore
- InfoLM

Future Work

Most of SOTA metrics are based on pretrained representations and lack of interpretability.

Introduce confidence scores.

Build news datasets.

Further Readings

Many Future works can be drawn from the works I presented !

Papers

Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing

Uncertainty-Aware Machine Translation Evaluation

Pushing the Right Buttons: Adversarial Evaluation of Quality Estimation

Better than Average: Paired Evaluation of NLP Systems

Researchers

Matt Post

https://scholar.google.com/citations?user=4w7LhxsAAAAJ&hl=en

Andre Martins

https://scholar.google.com/citations?user=mT7ppvwAAAAJ&hl=en

Lucia Specia

https://scholar.google.co.uk/citations?user=wVl_z8kAAAAJ&hl=en

Maxime Peyrard

https://scholar.google.fr/citations?user=RFMdKLMAAAAJ&hl=en

Other lines of research I am pursuing and you could be interested in!

Machine Learning & Security

- Learning Disentangled Representations
- Out of Distribution detection
- Miss-classification detection

Dialog Systems

- Learning Representations of Dialogs
- Multimodal Learning

Check out my scholar:

Acknowledgements

Computing Grant from GENCI.

This work would no have been possible without my co-authors

Guillaume Staerman, Chloé Clavel, Pavlo Mozharovskyi, Stephan Clémençon, Florence d'Alché-Buc, Pablo Piantanida

If you want to contact me:

colombo.pierre@gmail.com

Website

Twitter

Scholar

Automatique Evaluation of Natural Language Generation using Measures of Similarity

Thanks for your attention!

Conversational Agents

Definition

A conversational agent is any dialogue system that not only conducts natural language processing but also responds automatically using human language.

Natural Language Understanding (NLU)

The conversation agent has to understand the user.

Natural Language Generation (NLG)

The conversation agent has to produce a response.

Existing Methods

Edit Based

Snover et al. 2006

Operations

- Insertion (I)
- Deletion (D)
- Substitution (S).

tailor -> sailor (S)
sailor -> sailir (S)
sailir -> sailin (S)
sailin -> sailing (I)

Distance is 4 !

N-gram Based

Papineni et al. 2002

C : I like these very nice pies !

R : I like those cakes !

Unigrams

C : I like these very nice pies !

R : I like those cakes !

Bigrams

C : I like these very nice pies !

R : I like those cakes !

MoverScore & BaryScore

Embedding Based

Word Mover distance

Kusner et al. 2015

BertScore

Zhang et al. 2019

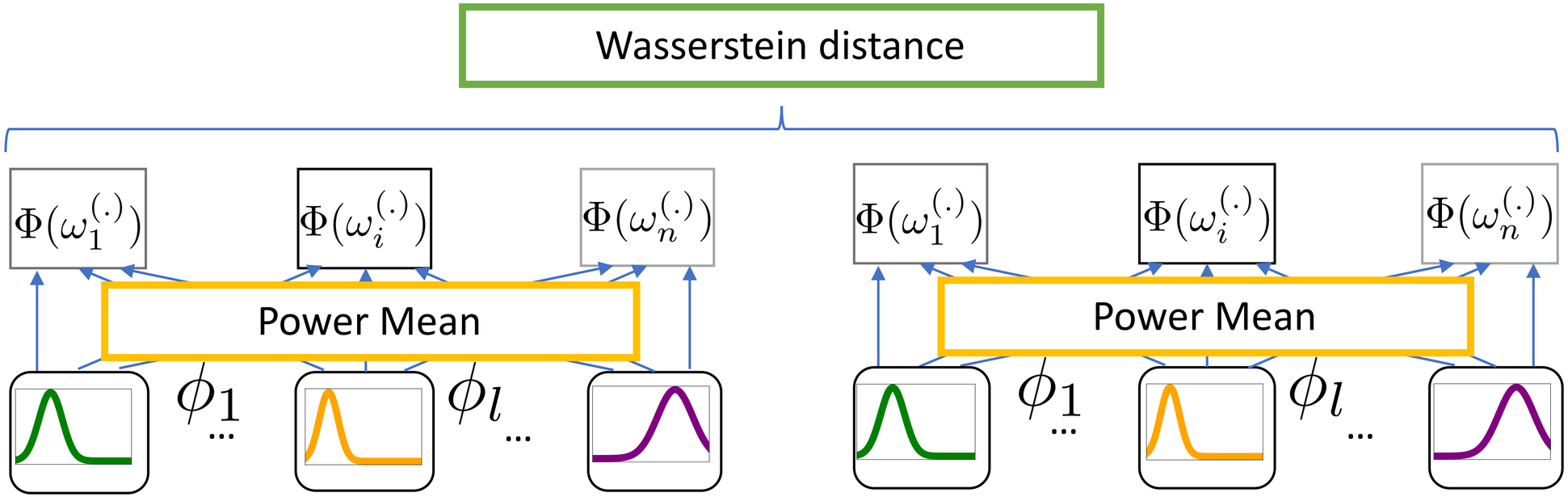
MoverScore

Zhao et al. 2019

Sentence Mover

Clark et al. 2019

MoverScore



R: The weather is cold today.

C: It's freezing today.

Advantage

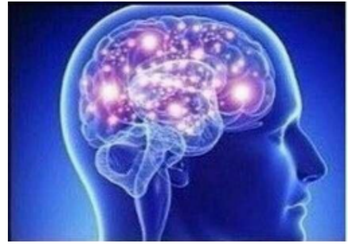
1. Deal with paraphrases
2. Include "semantic"
3. Use several layers

Limitations

1. Use arbitrary sequence of operation

(euclidean aggregation function Wasserstein distance)

Towards BaryScore



A novel metric called BaryScore

Previously

1 Take one layer

2 Do a series of operations (Wasserstein)

1 Take several layers

2 Aggregate using euclidean distance

3 Do a series of Operations (Wasserstein)

Best of all worlds

1 Take several layers

2 Aggregate using Wasserstein distance

3 Do a series of operation (Wasserstein)

BertScore

DepthScore

MoverScore

BaryScore