What are the best systems?

New Perspectives on NLP Benchmarking.

Nathan Noiry & Pierre Colombo

Datacraft 8. March 2022.

Takeaway of the presentation

Takeaway of the presentation

Classical AI pipeline:

Takeaway of the presentation

Classical AI pipeline:

Data collection

Takeaway of the presentation

Classical AI pipeline:

Data collection

Features extraction

Takeaway of the presentation

Classical AI pipeline:

Data collection

Features extraction

ML Model

Takeaway of the presentation

Classical AI pipeline:

Data collection

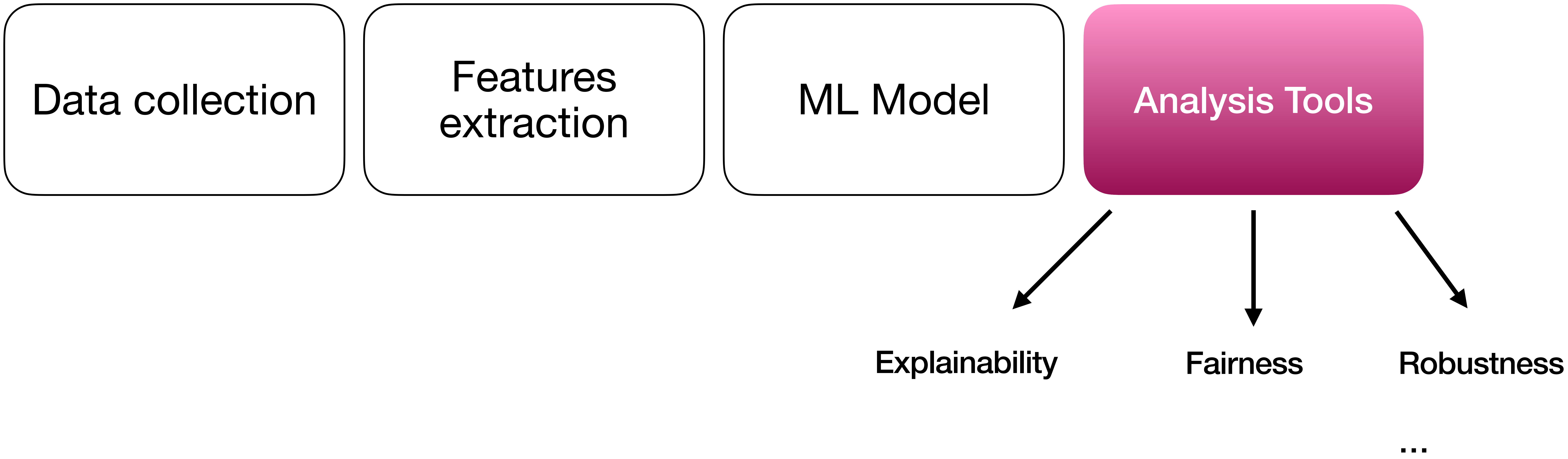
Features extraction

ML Model

Analysis Tools

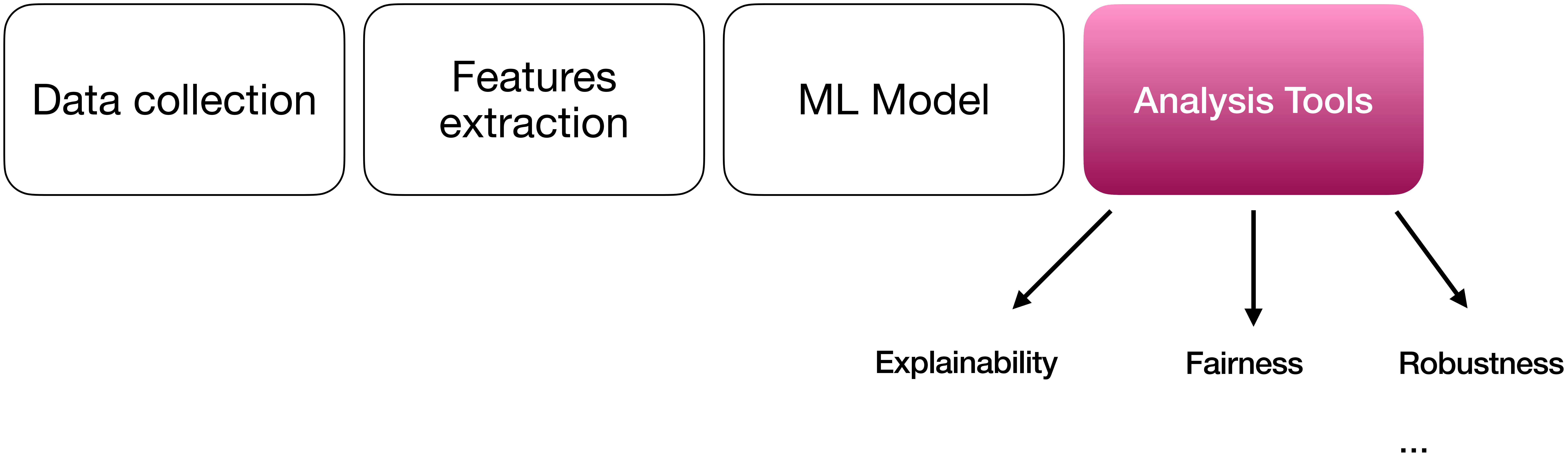
Takeaway of the presentation

Classical AI pipeline:



Takeaway of the presentation

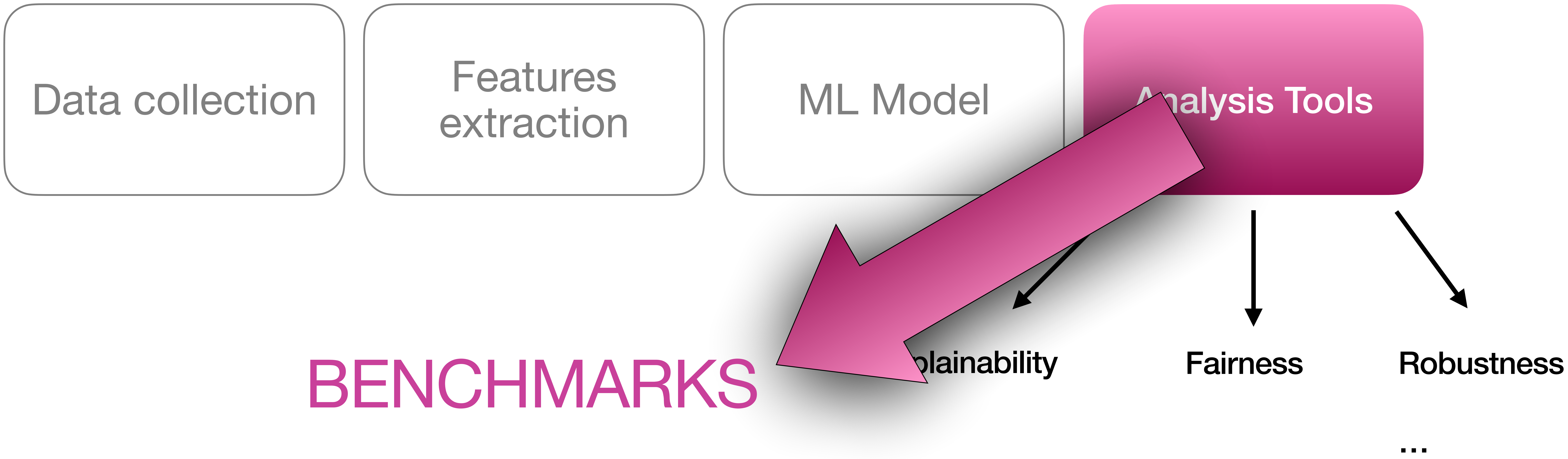
Classical AI pipeline:



Stop focusing on the models!

Takeaway of the presentation

Classical AI pipeline:



Stop focusing on the models!

Warmup

Warmup

What is a benchmark?

1. An ensemble of datasets
2. One or multiple metrics
3. A way to aggregate performances

Warmup

What is a benchmark?

1. An ensemble of datasets

2. One or multiple metrics

3. A way to aggregate performances

Why are benchmark vitals?

Research advances in Machine Learning are crucially fueled by *reliable evaluation procedures*

Outline

Outline

1. **How to evaluate Natural Language Generation?**

Outline

1. How to evaluate Natural Language Generation?

Outline

1. How to evaluate Natural Language Generation?

1.1 Context: problems, evaluation of automatic evaluation.

1.2 What are the main metrics to do reference based evaluation of NLG?

1.3 Reference based evaluation of NLG using embedding based metrics.

1.4 Beyond embedding based metrics.

2. How to aggregate several metrics?

Outline

1. How to evaluate Natural Language Generation?

1.1 Context: problems, evaluation of automatic evaluation.

1.2 What are the main metrics to do reference based evaluation of NLG?

1.3 Reference based evaluation of NLG using embedding based metrics.

1.4 Beyond embedding based metrics.

2. How to aggregate several metrics?

1.1 Framework

1.2 Task Level Aggregation

1.3 Instance Level Aggregation

Outline

1. How to evaluate Natural Language Generation?

1.1 Context: problems, evaluation of automatic evaluation.

1.2 What are the main metrics to do reference based evaluation of NLG?

1.3 Reference based evaluation of NLG using embedding based metrics.

1.4 Beyond embedding based metrics.

2. How to aggregate several metrics?

1.1 Framework

1.2 Task Level Aggregation

1.3 Instance Level Aggregation

3. Conclusions

1. How to evaluate Natural Language Generation?

1.1 Context: problems, evaluation of automatic evaluation.

1.2 What are the main metrics to do reference based evaluation of NLG?

1.3 Reference based evaluation of NLG using embedding based metrics.

1.4 Beyond embedding based metrics.

1. How to evaluate Natural Language Generation?

1.1 Context: problems, evaluation of automatic evaluation.

1.2 What are the main metrics to do reference based evaluation of NLG?

1.3 Reference based evaluation of NLG using embedding based metrics.

1.4 Beyond embedding based metrics.

When do we need automatic evaluation?

When do we need automatic evaluation?

1. Debug NLG systems without annotators.

When do we need automatic evaluation?

1. Debug NLG systems without annotators.

2. Improve learning of systems by deriving new losses.

When do we need automatic evaluation?

1. Debug NLG systems without annotators.

2. Improve learning of systems by deriving new losses.

3. Compare different systems.

When do we need automatic evaluation?

1. Debug NLG systems without annotators.

2. Improve learning of systems by deriving new losses.

3. Compare different systems.

Why do we need human evaluation?

When do we need automatic evaluation?

1. Debug NLG systems without annotators.

2. Improve learning of systems by deriving new losses.

3. Compare different systems.

Why do we need human evaluation?

1. Cheap: compared to human evaluation.

When do we need automatic evaluation?

1. Debug NLG systems without annotators.

2. Improve learning of systems by deriving new losses.

3. Compare different systems.

Why do we need human evaluation?

1. Cheap: compared to human evaluation.

2. Fast: you can label “instantaneously”.

When do we need automatic evaluation?

1. Debug NLG systems without annotators.

2. Improve learning of systems by deriving new losses.

3. Compare different systems.

Karpinska et al. 2021

Why do we need human evaluation?

1. Cheap: compared to human evaluation.

2. Fast: you can label “instantaneously”.

3. Reproductible: two sentences always get the same score.

When do we need automatic evaluation?

1. Debug NLG systems without annotators.

2. Improve learning of systems by deriving new losses.

3. Compare different systems.

Karpinska et al. 2021

Why do we need human evaluation?

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).

Let’s formalize the problem of Automatic Evaluation

Let's formalize the problem of Automatic Evaluation

S_1: The weather is cold today.

S_2: It is freezing today

0.8

Let’s formalize the problem of Automatic Evaluation

S_1: The weather is cold today.

S_2: It is freezing today

→ 0.8

S_1: I like those cats.

S_2: It is freezing today

→ 0.1

Let's formalize the problem of Automatic Evaluation

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

Let's formalize the problem of Automatic Evaluation

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

We want to build a metric m

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

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

Let's formalize the problem of Automatic Evaluation

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

We want to build 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?

Let's formalize the problem of Automatic Evaluation

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

We want to build 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

Reference based vs reference free evaluation

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

Reference based vs reference free evaluation

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

Scenario 1: Let's assume we have a reference

Reference based

Reference based vs reference free evaluation

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

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

Scenario 1: Let's assume we have a reference
Reference based

System

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

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

Reference based vs reference free evaluation

Scenario 1: Let's assume we have a reference
Reference based

System

Human

Reference based vs reference free evaluation

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

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

Scenario 1: Let's assume we have a reference
Reference based

System

I like running outside !

Human

Reference based vs reference free evaluation

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

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

Scenario 1: Let's assume we have a reference
Reference based

I like running outside !

S_1 **J'aime courir à l'extérieur !**

System

Human

Reference based vs reference free evaluation

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

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

Scenario 1: Let's assume we have a reference
Reference based

S_1 J'aime courir à l'extérieur !

System

I like running outside !

S_2 J'apprécie courir dehors !

Human

Reference based vs reference free evaluation

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

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

Scenario 1: Let's assume we have a reference
Reference based

I like running outside !

System

S_1 J'aime courir à l'extérieur !

Human

S_2 J'apprécie courir dehors !

Scenario 2: we have no reference.
Reference free

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

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

Reference based vs reference free evaluation

Scenario 1: Let's assume we have a reference
Reference based

I like running outside !

System

S_1 J'aime courir à l'extérieur !

Human

S_2 J'apprécie courir dehors !

Scenario 2: we have no reference.
Reference free

S_1 I like running outside !

Reference based vs reference free evaluation

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

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

Scenario 1: Let's assume we have a reference
Reference based

I like running outside !

System

S_1 J'aime courir à l'extérieur !

Human

S_2 J'apprécie courir dehors !

Scenario 2: we have no reference.
Reference free

S_1 I like running outside !

S_2 J'aime courir à l'extérieur !

Reference based vs reference free evaluation

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

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

Scenario 1: Let's assume we have a reference
Reference based

I like running outside !

System

S_1 J'aime courir à l'extérieur !

Human

S_2 J'apprécie courir dehors !

Scenario 2: we have no reference.
Reference free

S_1 I like running outside !

S_2 J'aime courir à l'extérieur !

1. How to evaluate Natural Language Generation?

1.1 Context: problems, evaluation of automatic evaluation.

1.2 What are the main metrics to do reference based evaluation of NLG?

1.3 Reference based evaluation of NLG using embedding based metrics.

1.4 Beyond embedding based metrics.

Existing Metrics for Reference Based NLG

Existing Metrics for Reference Based NLG

Goal

R: **The weather is cold today.**

C: **It is freezing today**

0.8

Similar

R: **I like those cats.**

C: **It is freezing today**

0.1

Dissimilar

Existing Metrics for Reference Based NLG

Goal

R: The weather is cold today.
C: It is freezing today

0.8 Similar

R: I like those cats.
C: It is freezing today

0.1 Dissimilar

Edit Based

Existing Metrics for Reference Based NLG

Goal

R: The weather is cold today.
C: It is freezing today → 0.8 Similar

R: I like those cats.
C: It is freezing today → 0.1 Dissimilar

Edit Based

N-gram Based

Existing Metrics for Reference Based NLG

Goal

R: The weather is cold today.
C: It is freezing today

0.8 Similar

R: I like those cats.
C: It is freezing today

0.1 Dissimilar

Edit Based

N-gram Based

Embedding Based

Existing Methods

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 !

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 !

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

1. How to evaluate Natural Language Generation?

1.1 Context: problems, evaluation of automatic evaluation.

1.2 What are the main metrics to do reference based evaluation of NLG?

1.3 Reference based evaluation of NLG using embedding based metrics.

1.4 Beyond embedding based metrics.

Embedding Based

Embedding Based

Intuition

R: The weather is cold today.

C: It is freezing today

→ 0.8

Embedding Based

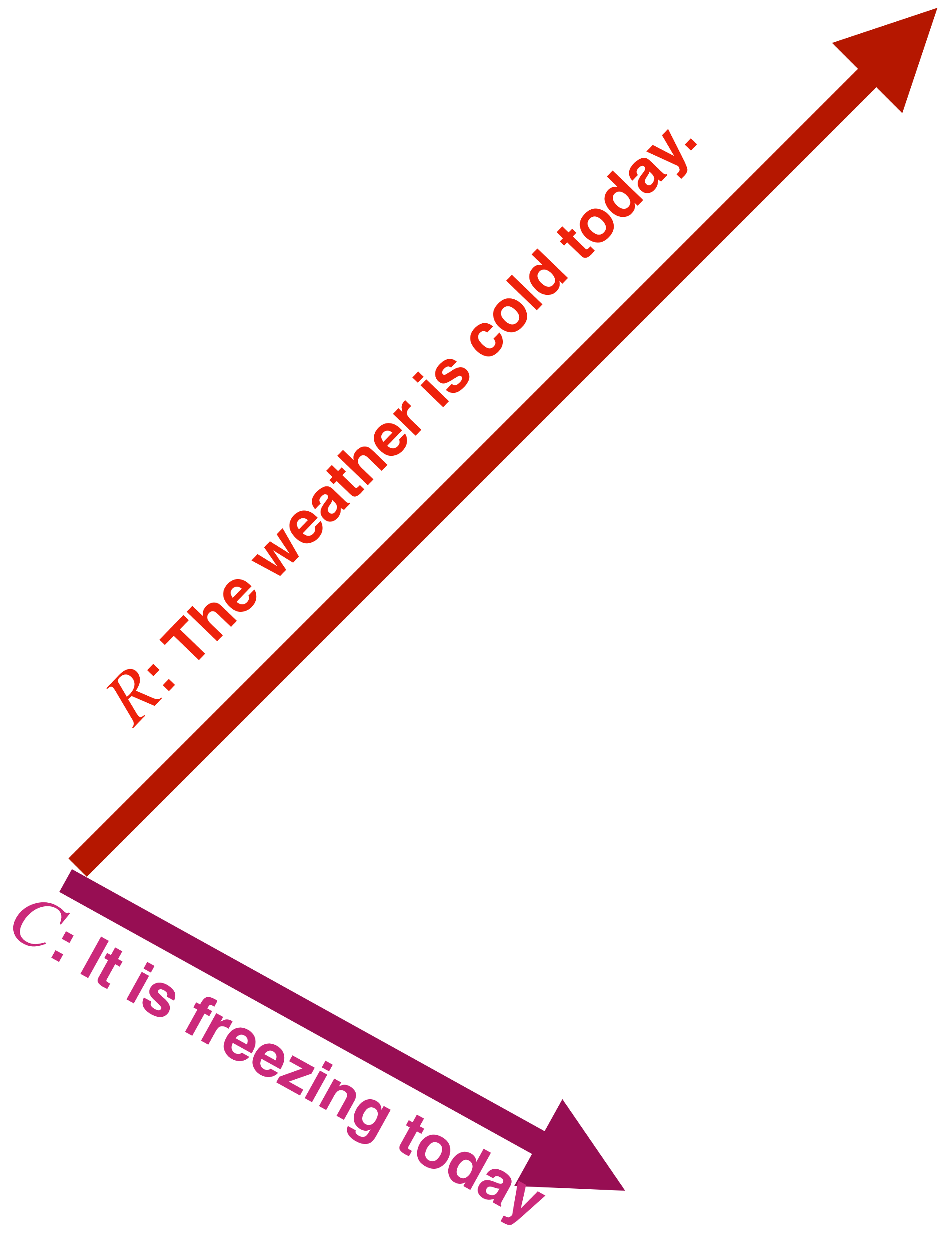
Intuition

R: The weather is cold today.

C: It is freezing today

0.8

1. Choose your embedding



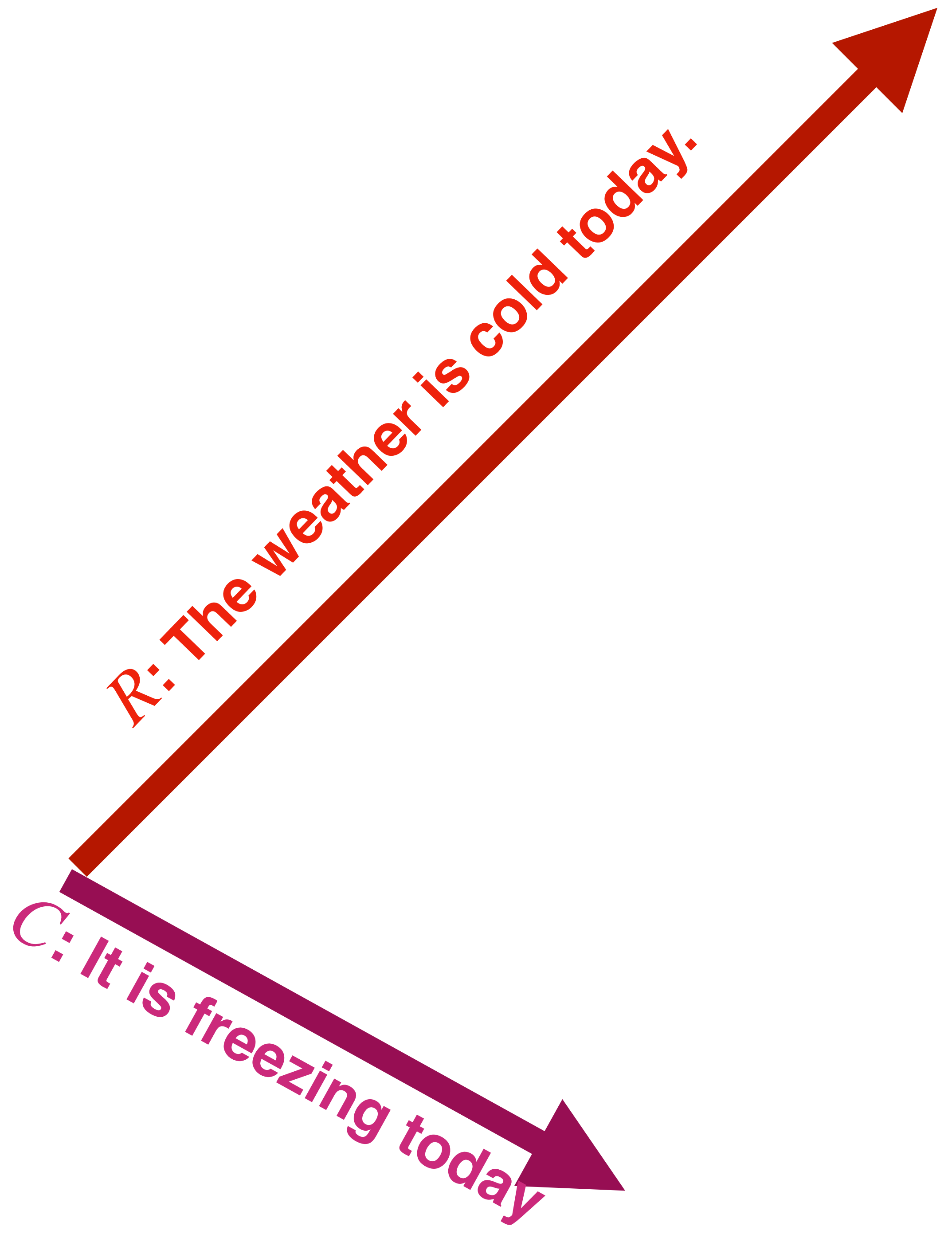
Embedding Based

Intuition

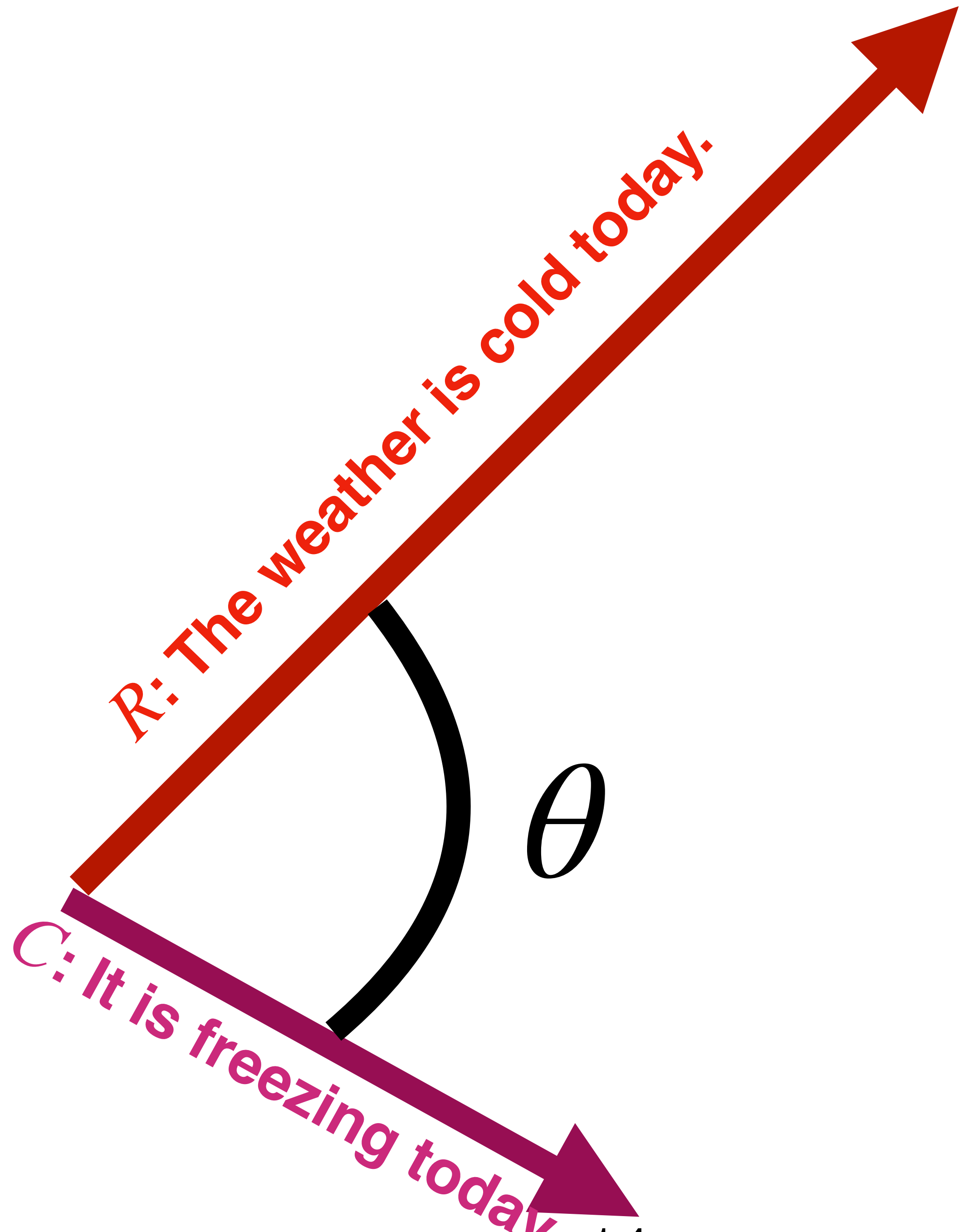
R: The weather is cold today.

C: It is freezing today

1. Choose your embedding



2. Choose a similarity function



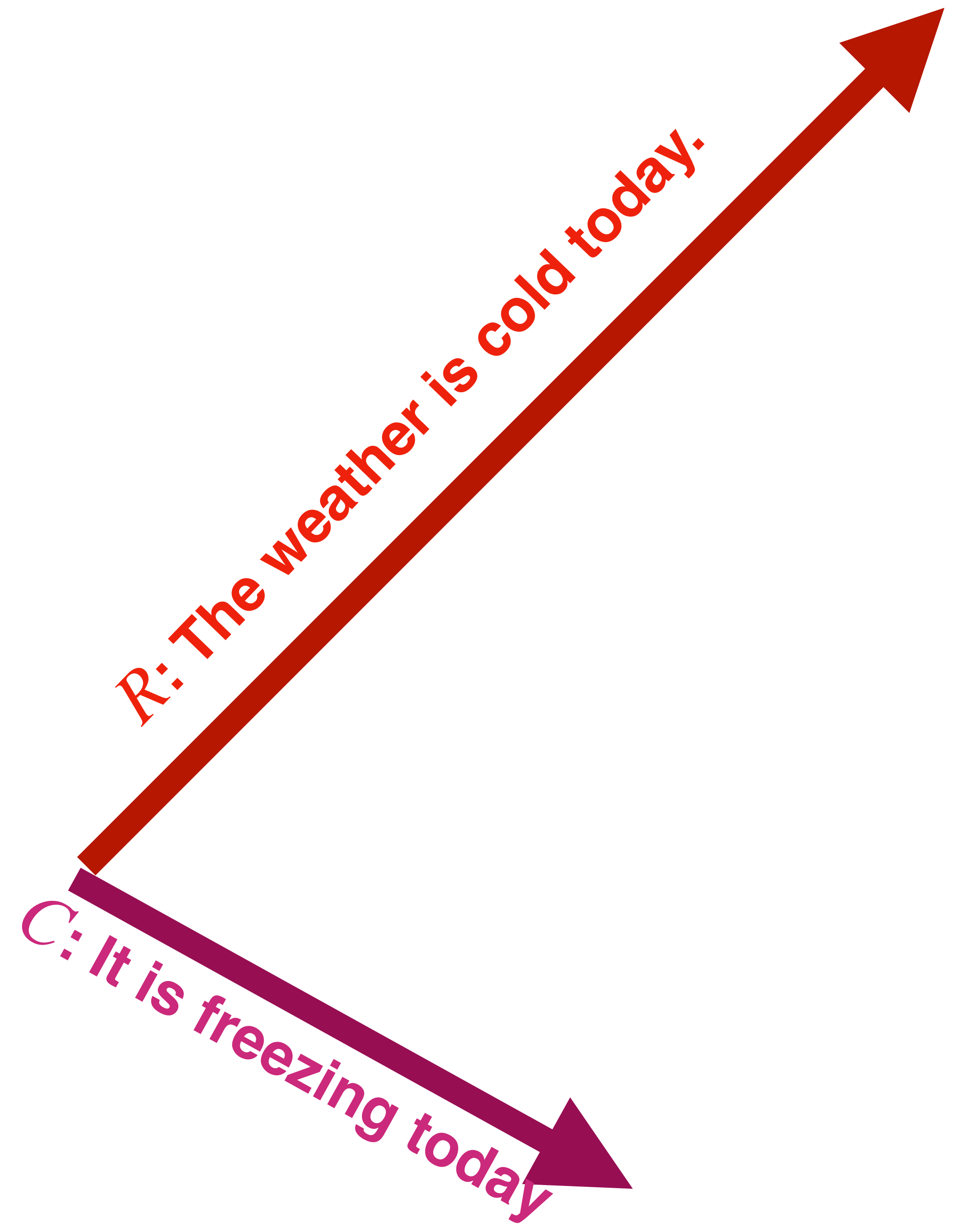
Embedding Based

Intuition

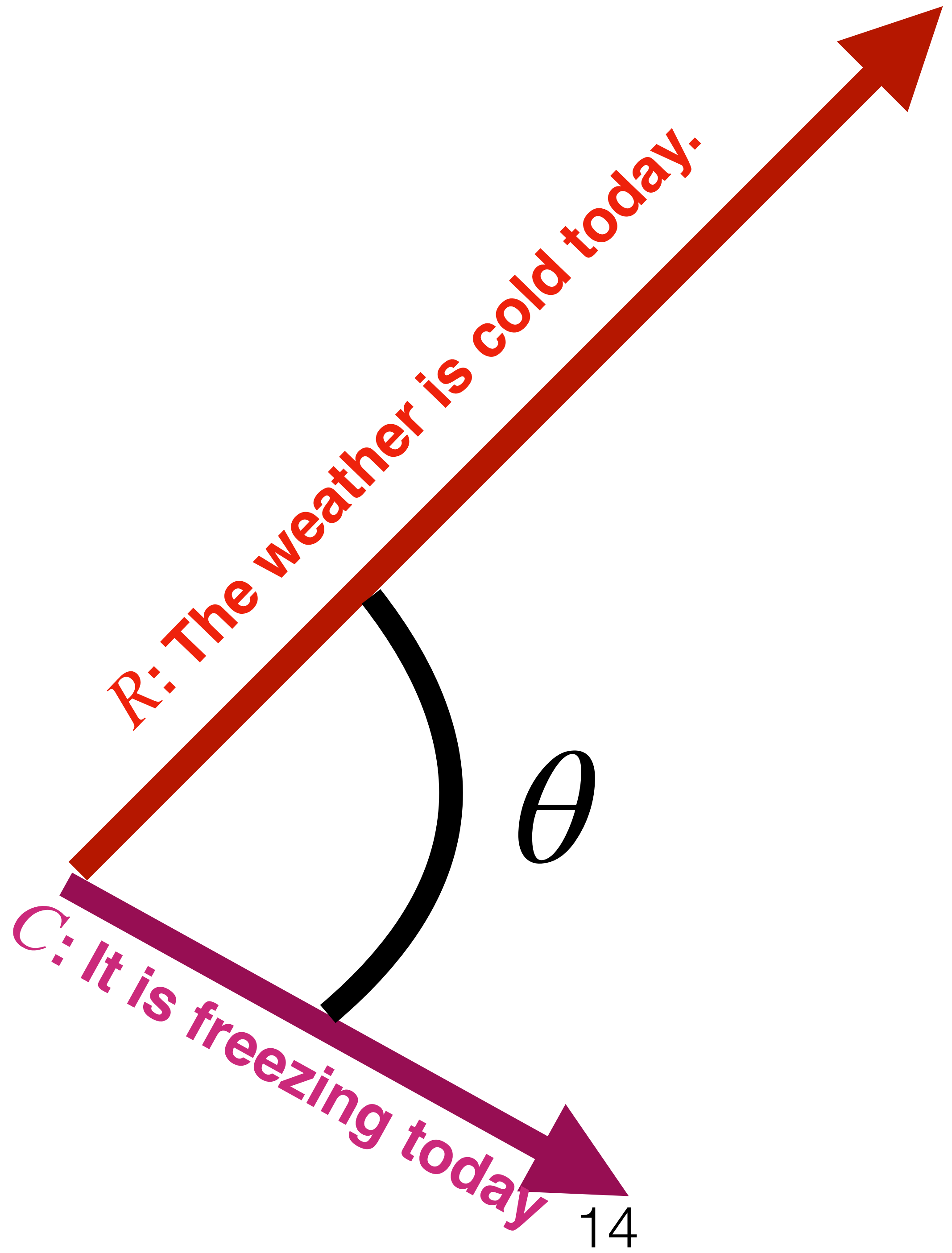
R: The weather is cold today.

C: It is freezing today

1. Choose your embedding



2. Choose a similarity function



Advantage

1. Deal with paraphrases
2. Include "semantic"

Embedding Based

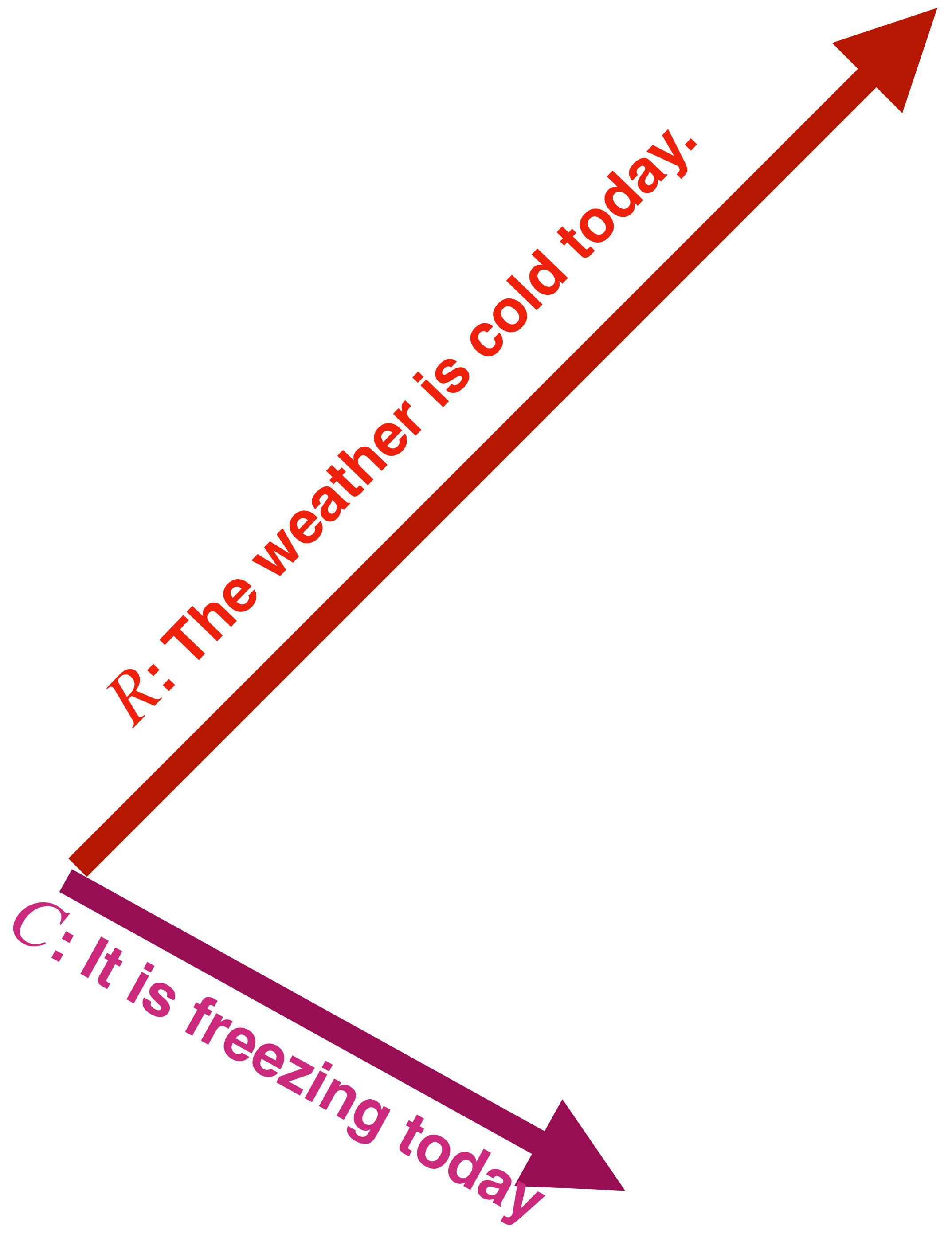
Intuition

R: The weather is cold today.

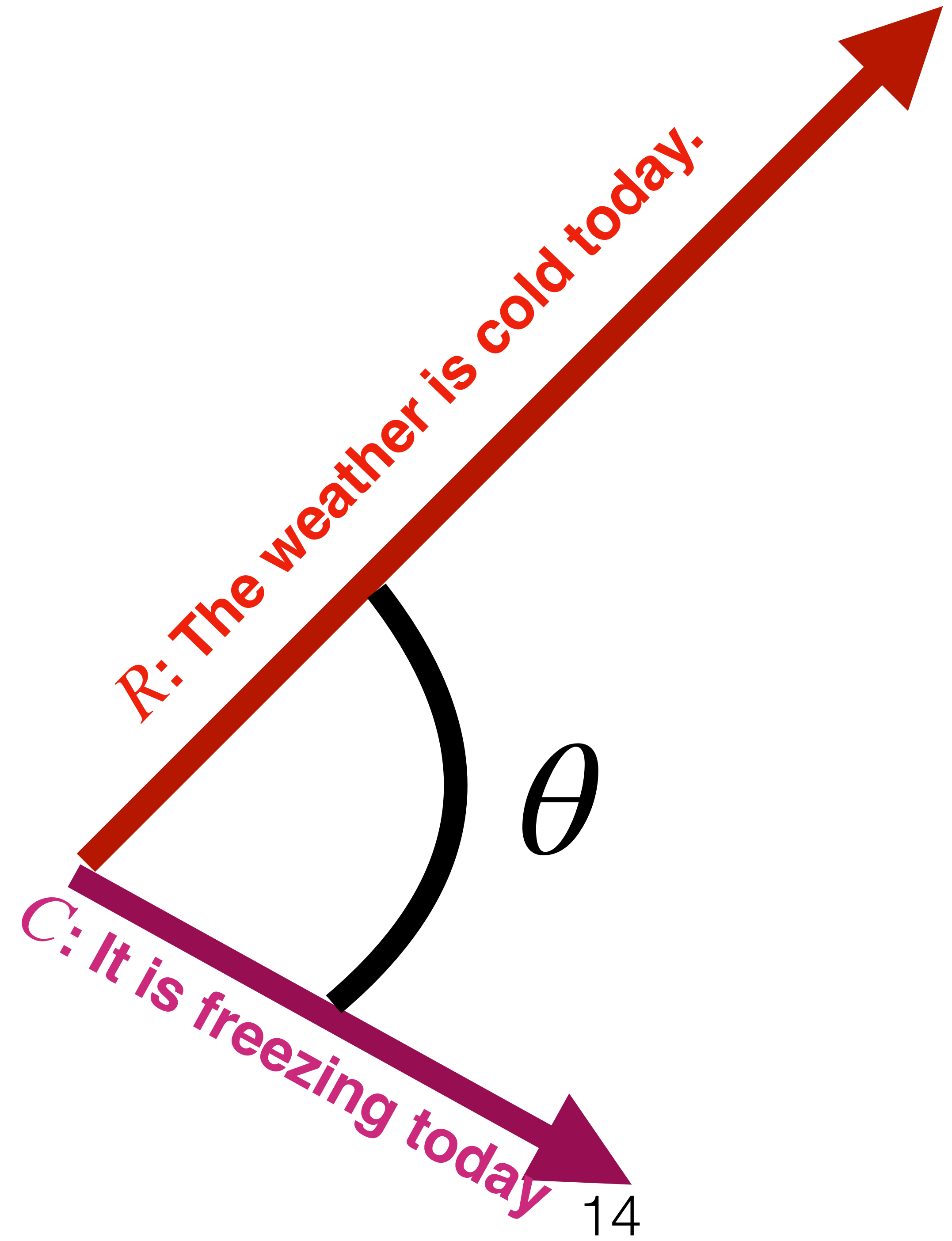
C: It is freezing today

0.8

1. Choose your embedding



2. Choose a similarity function



Advantage

1. Deal with paraphrases
2. Include "semantic"

Limitation

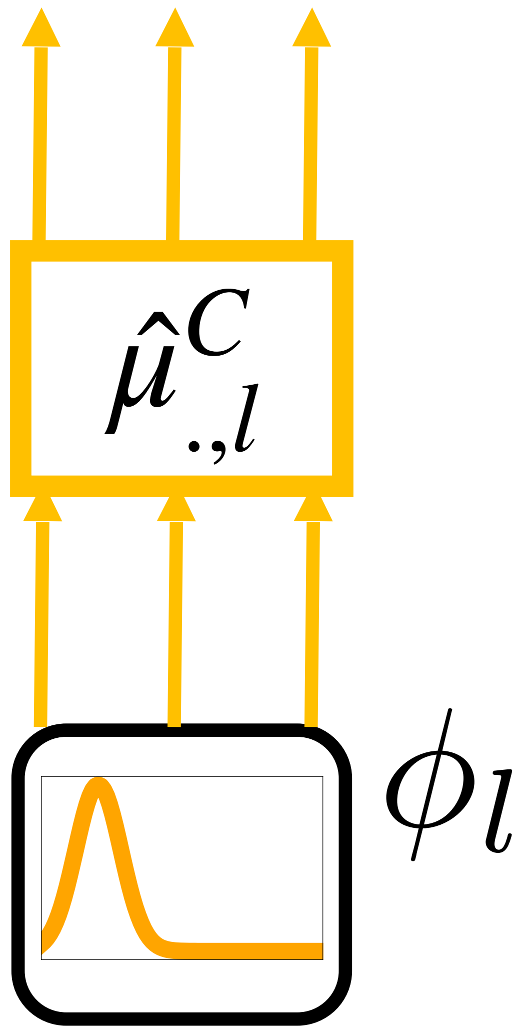
1. Not interpretable

BertScore

Zhang et al. 2019

BertScore

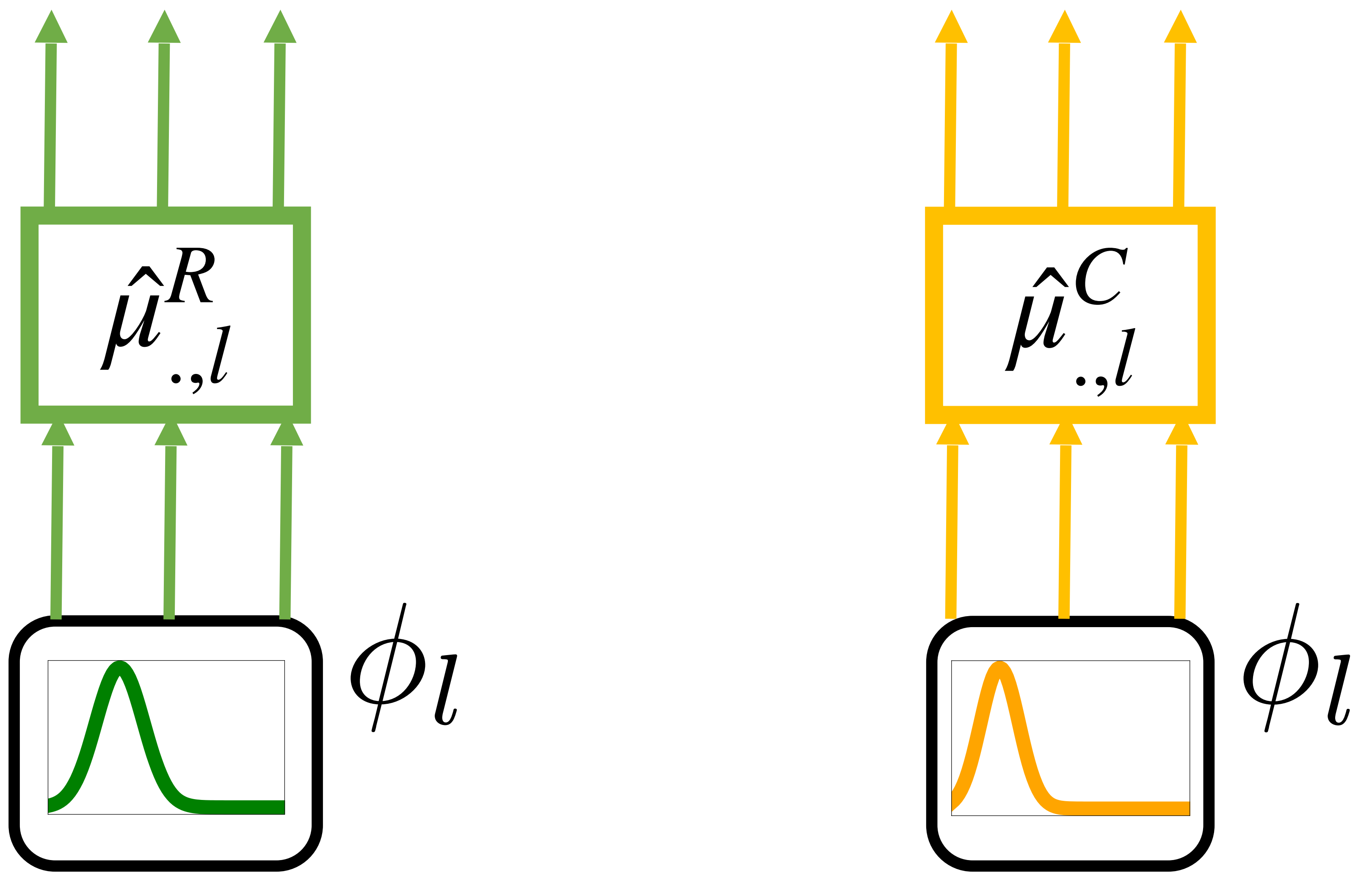
Zhang et al. 2019



C: It is freezing this morning

BertScore

Zhang et al. 2019

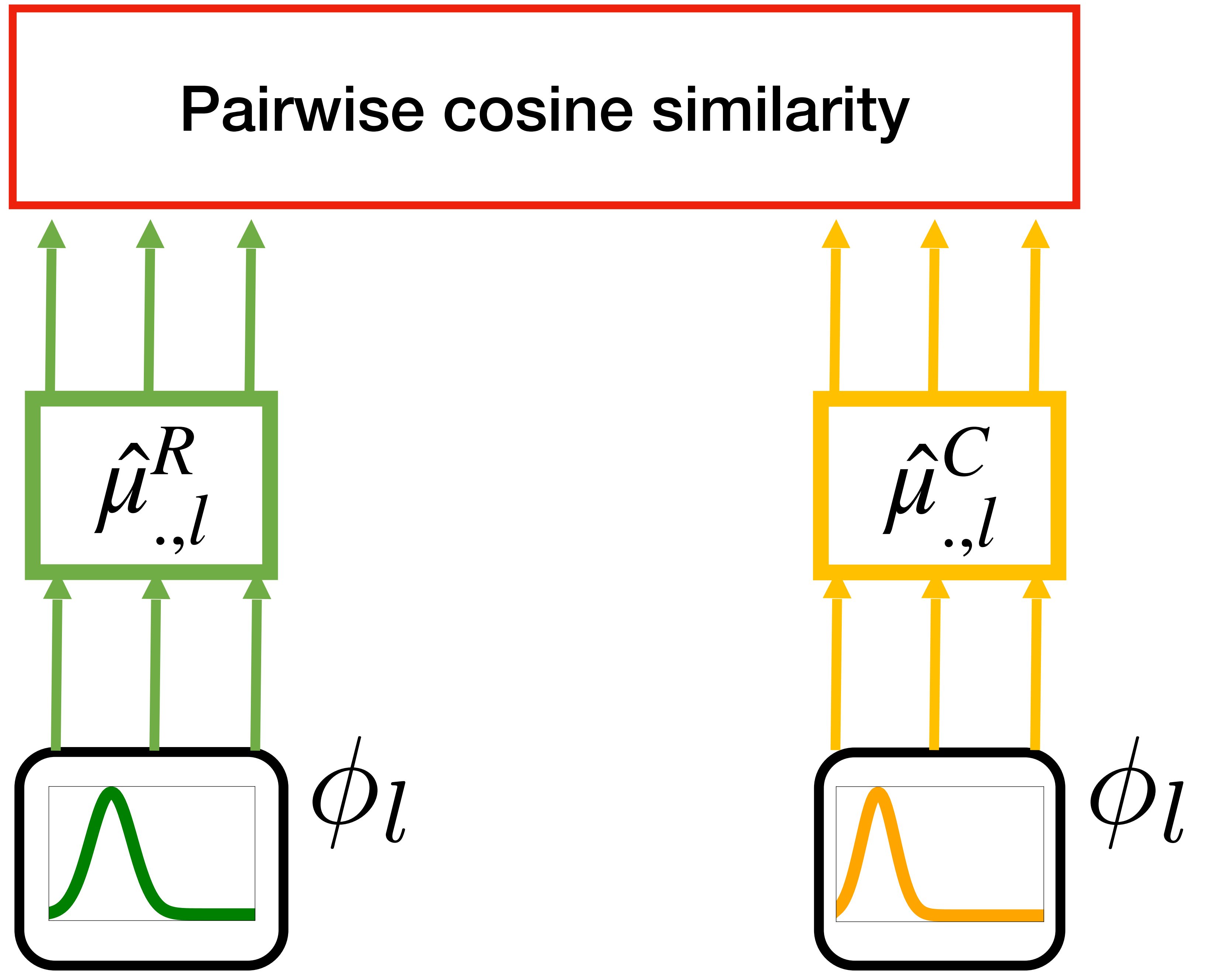


R: The weather is cold today

C: It is freezing this morning

BertScore

Zhang et al. 2019

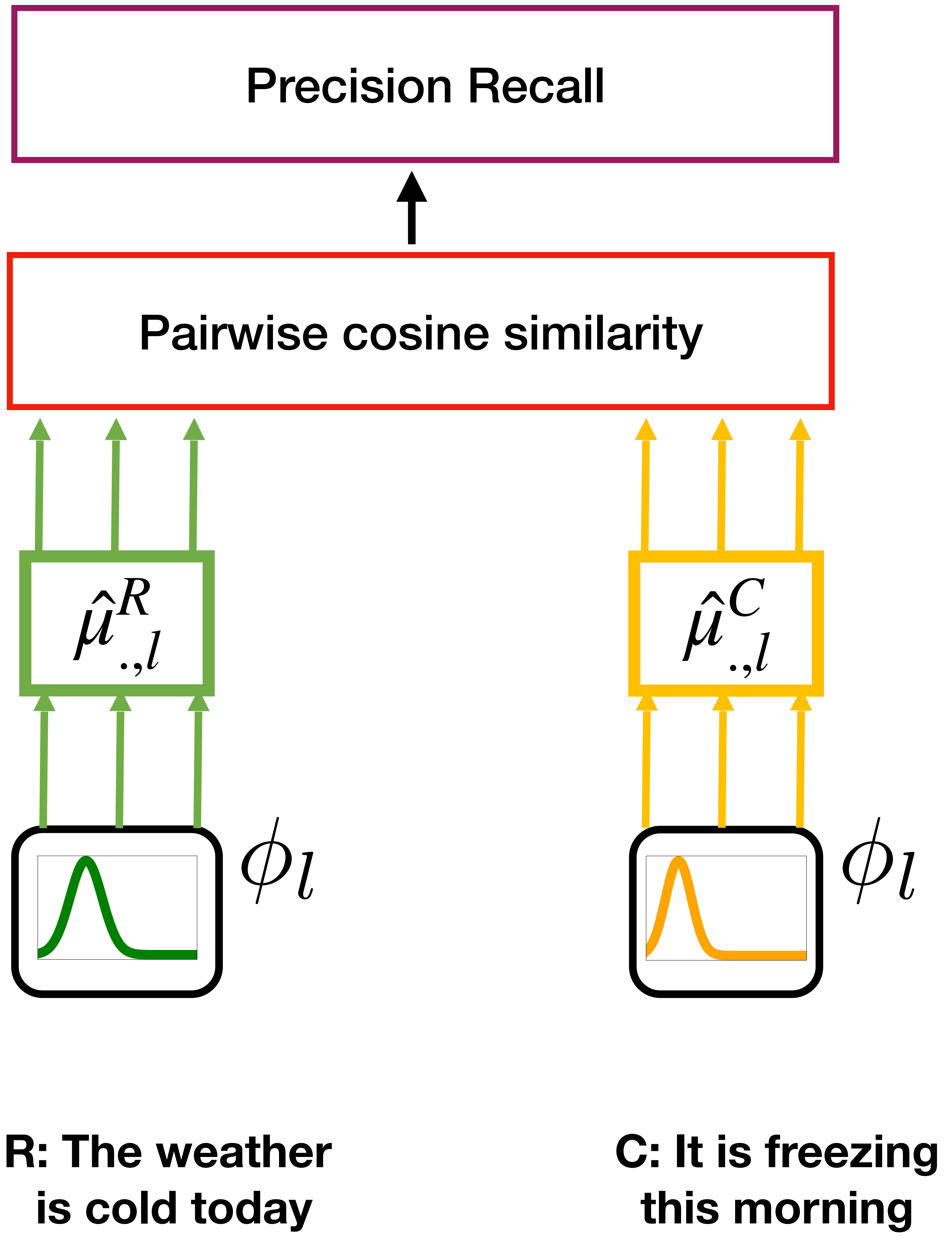


R: The weather is cold today

C: It is freezing this morning

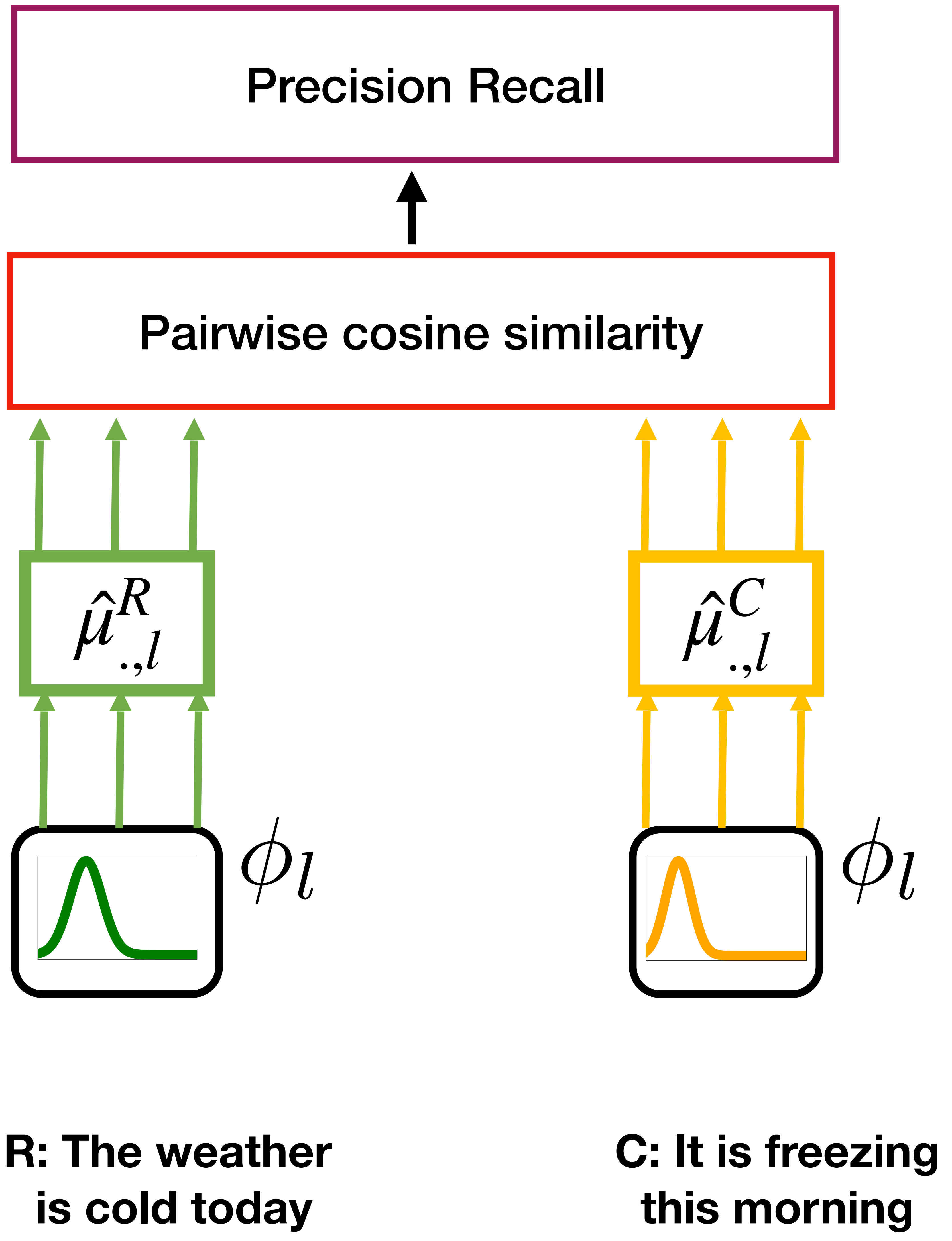
BertScore

Zhang et al. 2019



BertScore

Zhang et al. 2019

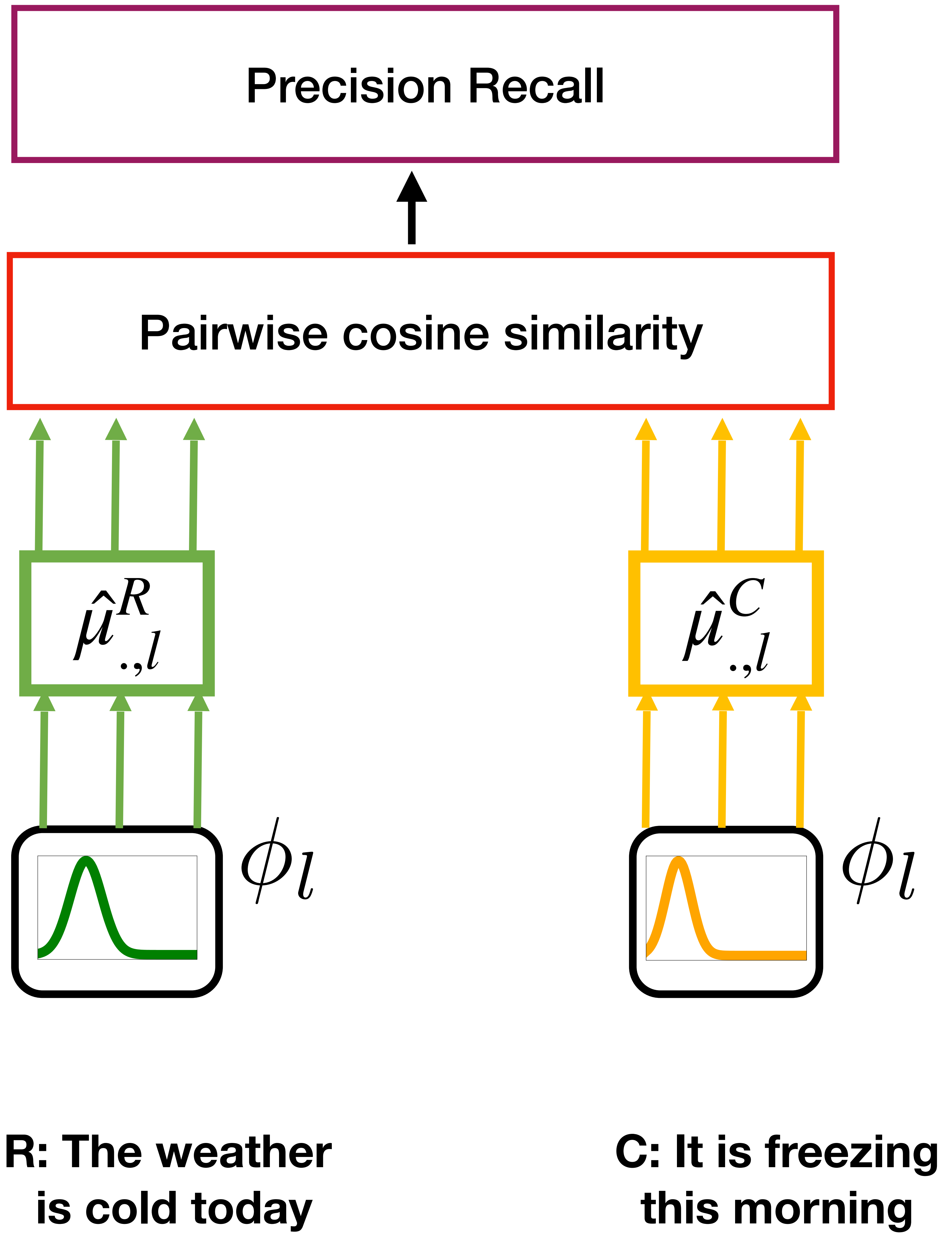


Advantage

1. Deal with paraphrases
2. Include "semantic"

BertScore

Zhang et al. 2019



Advantage

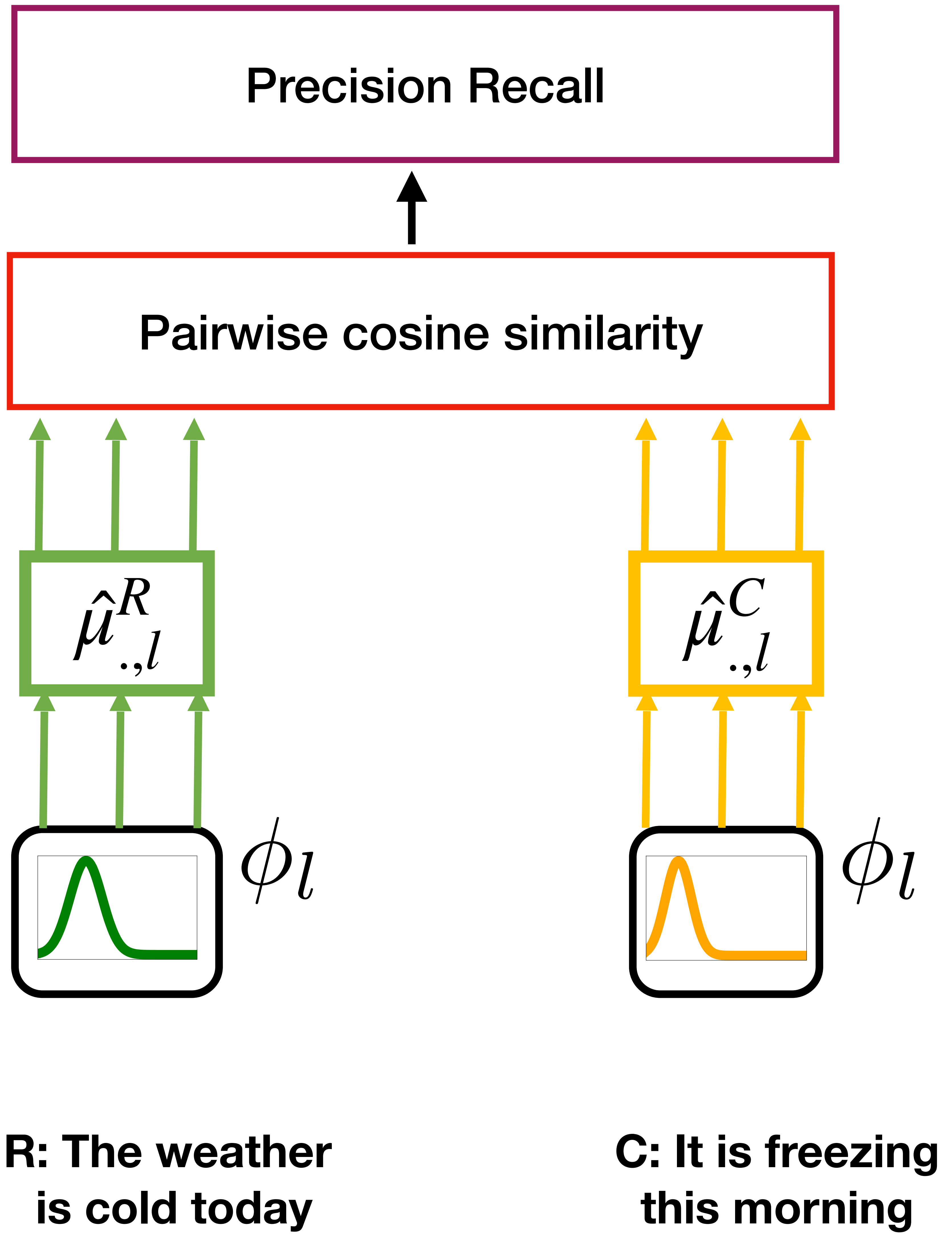
1. Deal with paraphrases
2. Include "semantic"

Limitations

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

BertScore

Zhang et al. 2019



Advantage

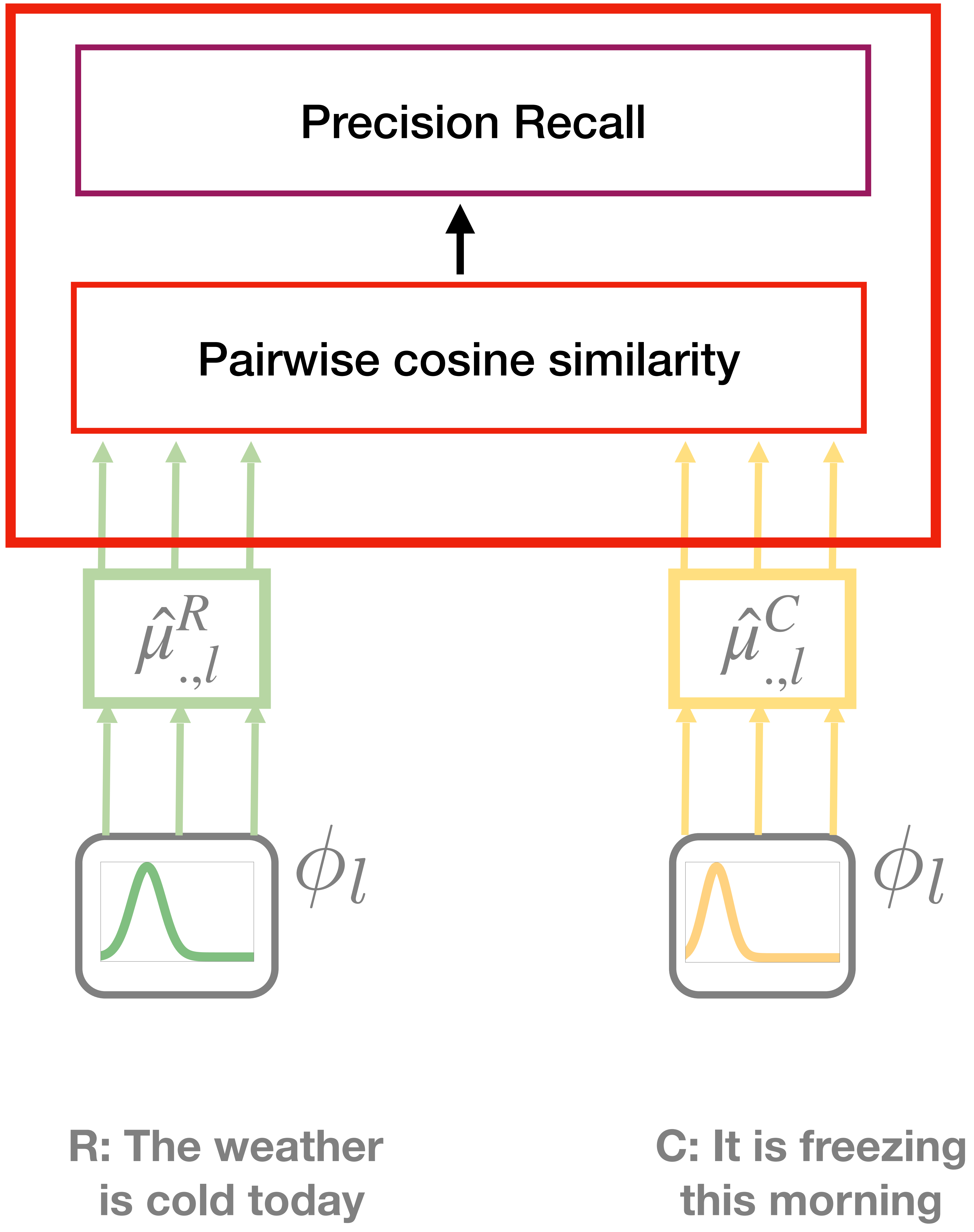
1. Deal with paraphrases
2. Include "semantic"

Limitations

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

BertScore

Zhang et al. 2019



This is a distance between two empirical distributions !

Advantage

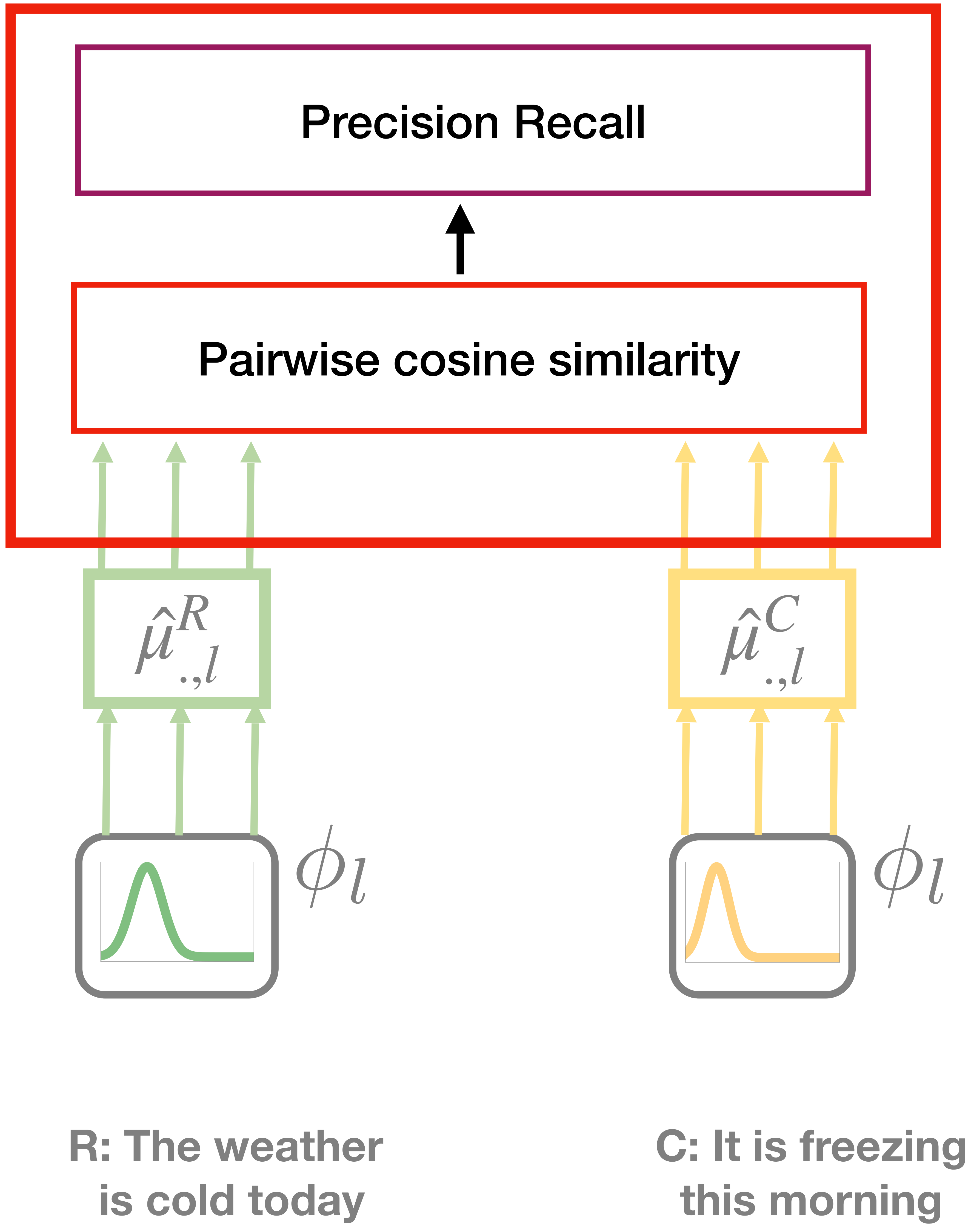
1. Deal with paraphrases
2. Include "semantic"

Limitations

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

BertScore

Zhang et al. 2019



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

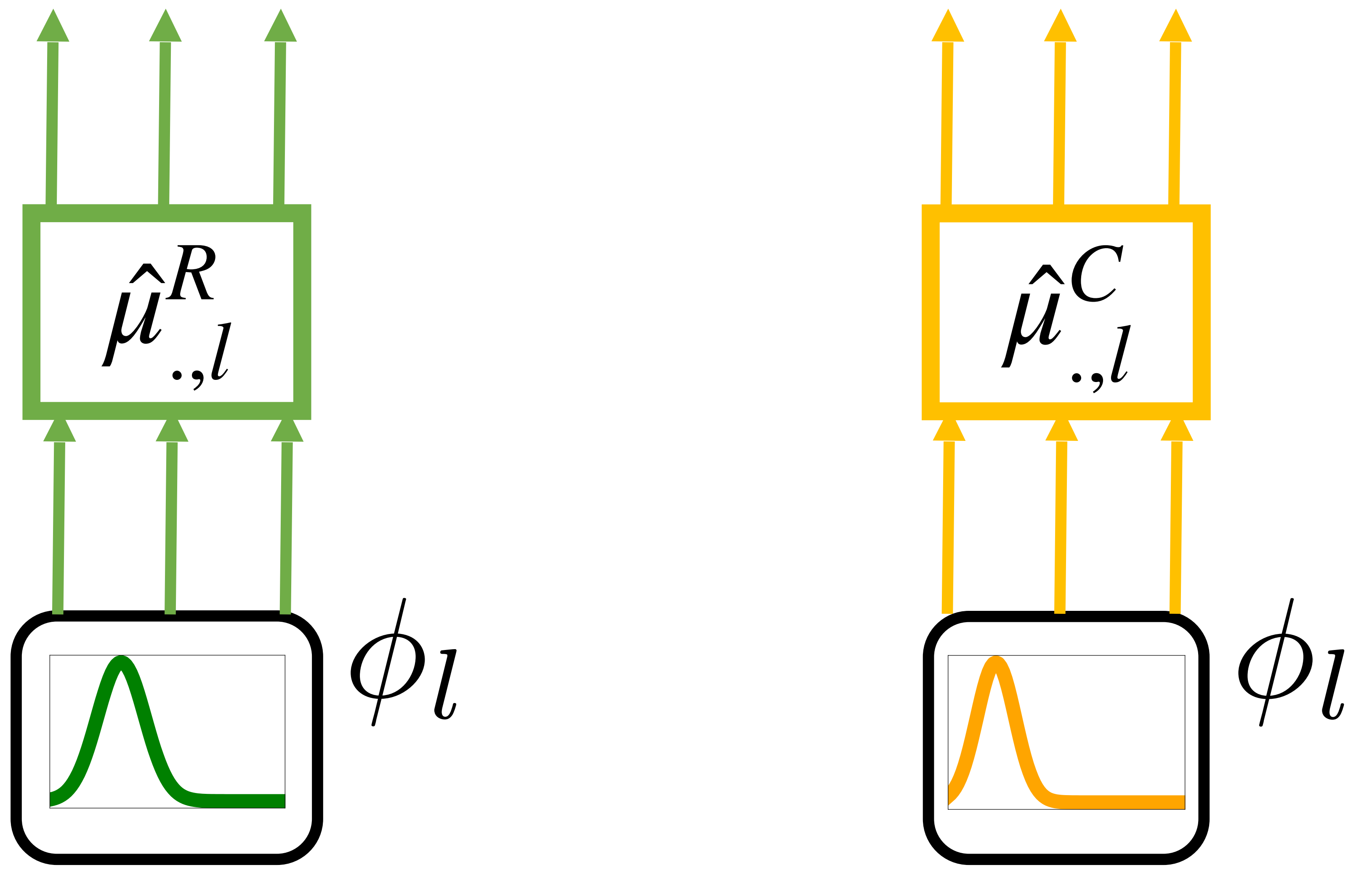
Still not interpretable

DepthScore

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

DepthScore

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



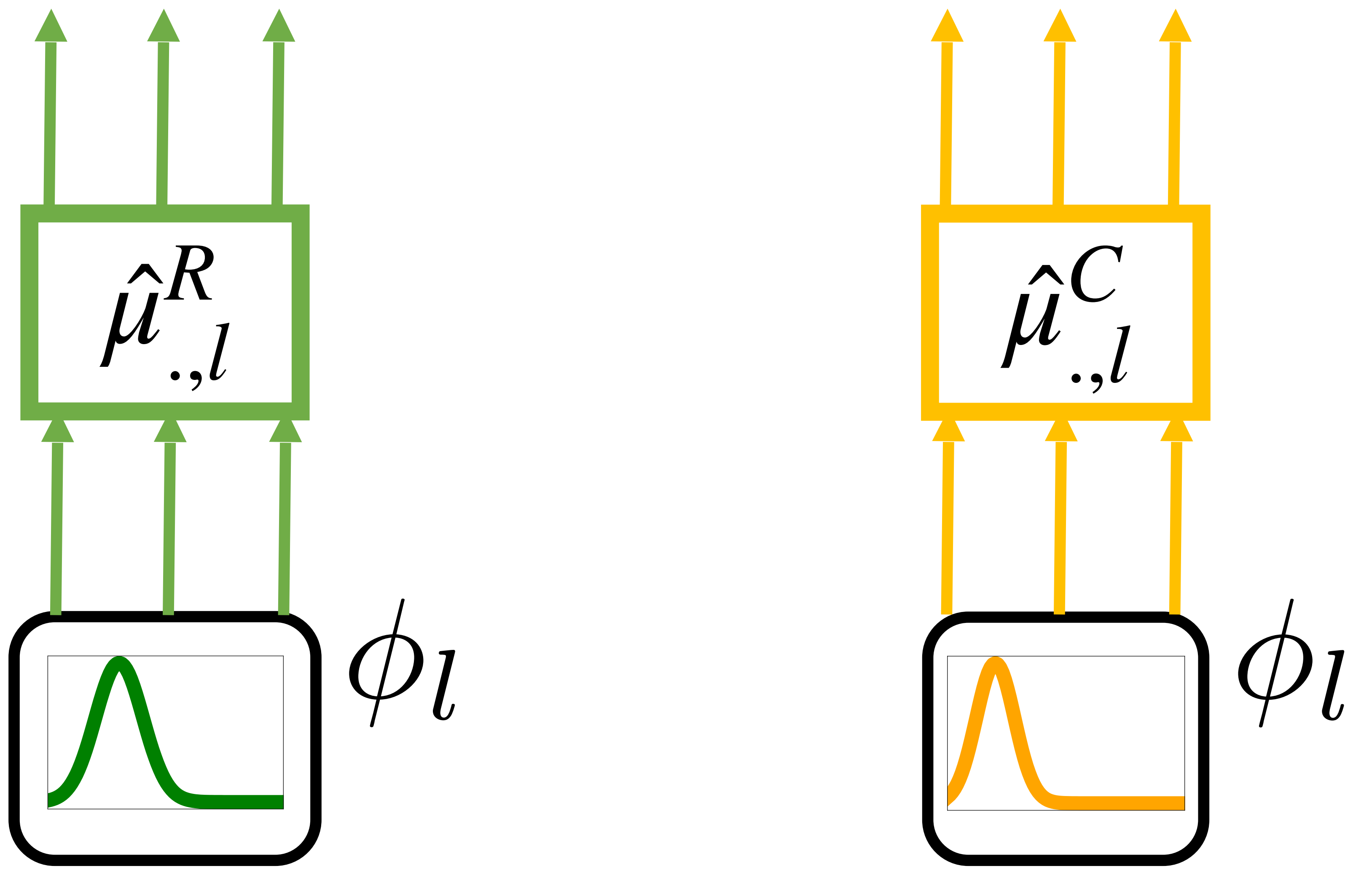
R: The weather is cold today

C: It is freezing this morning

DepthScore

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

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

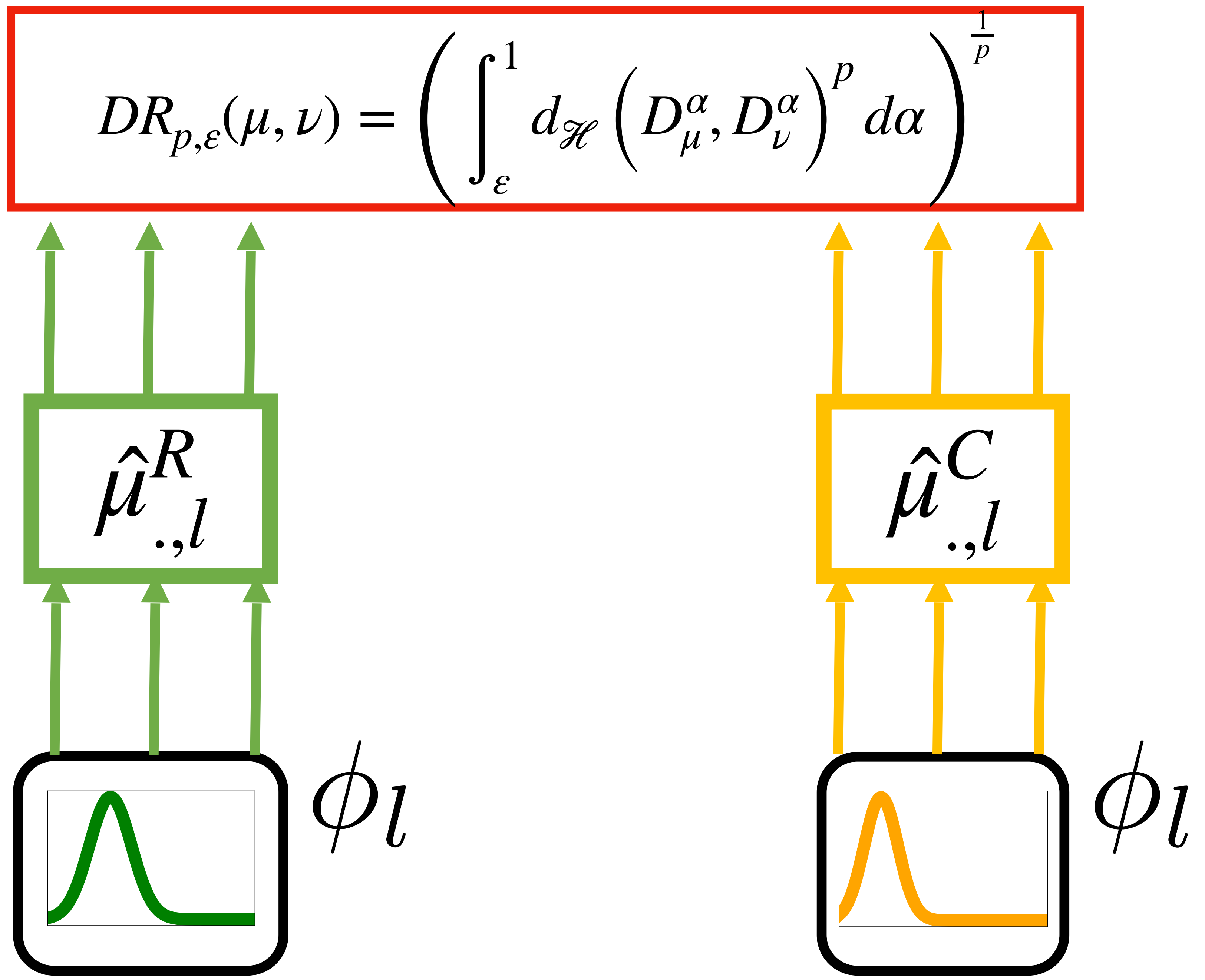


R: The weather is cold today

C: It is freezing this morning

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

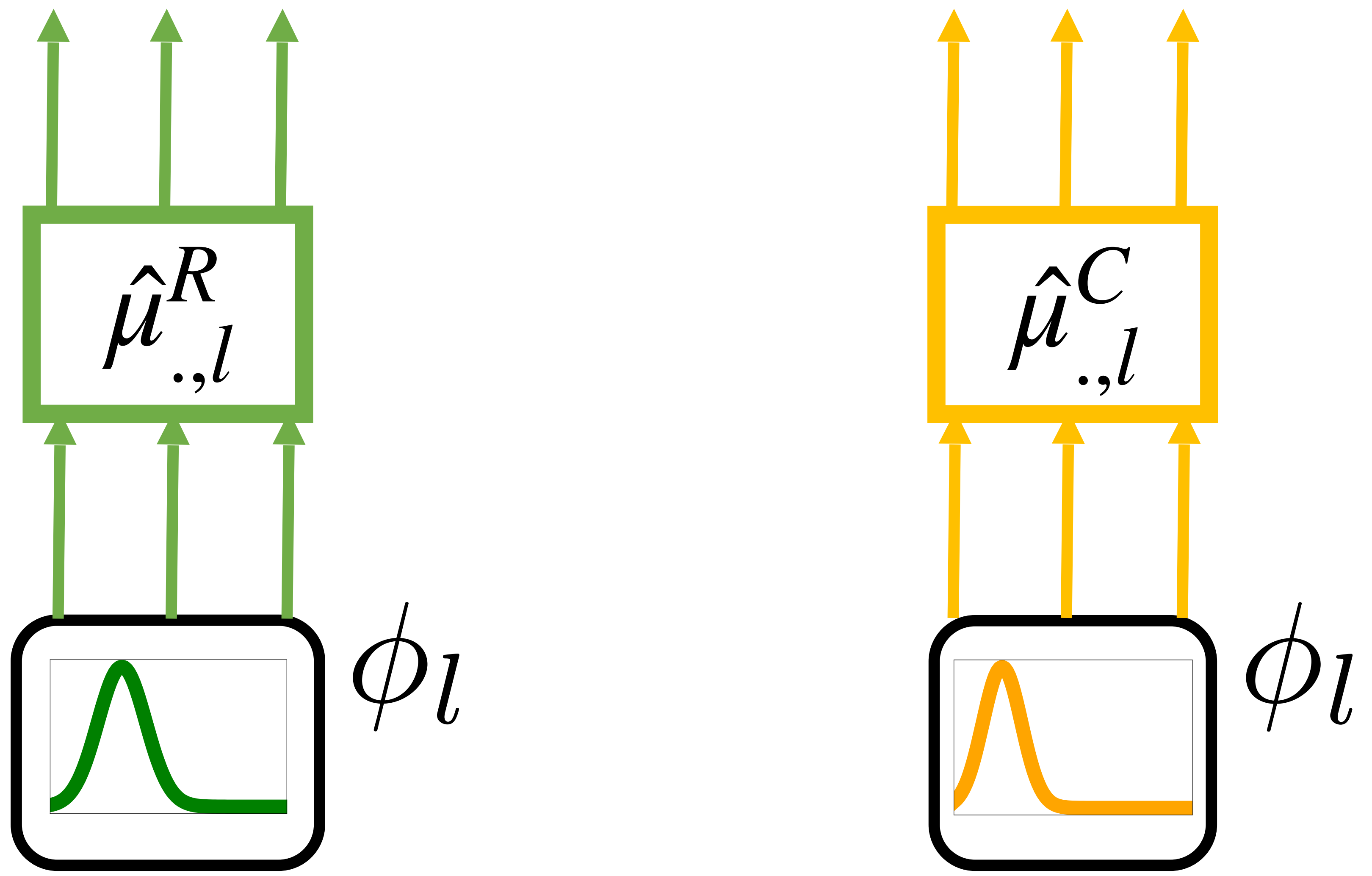
Advantage

1. **Deal with paraphrases**
2. **Include "semantic"**

DepthScore

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

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



R: The weather is cold today

C: It is freezing this morning

Advantage

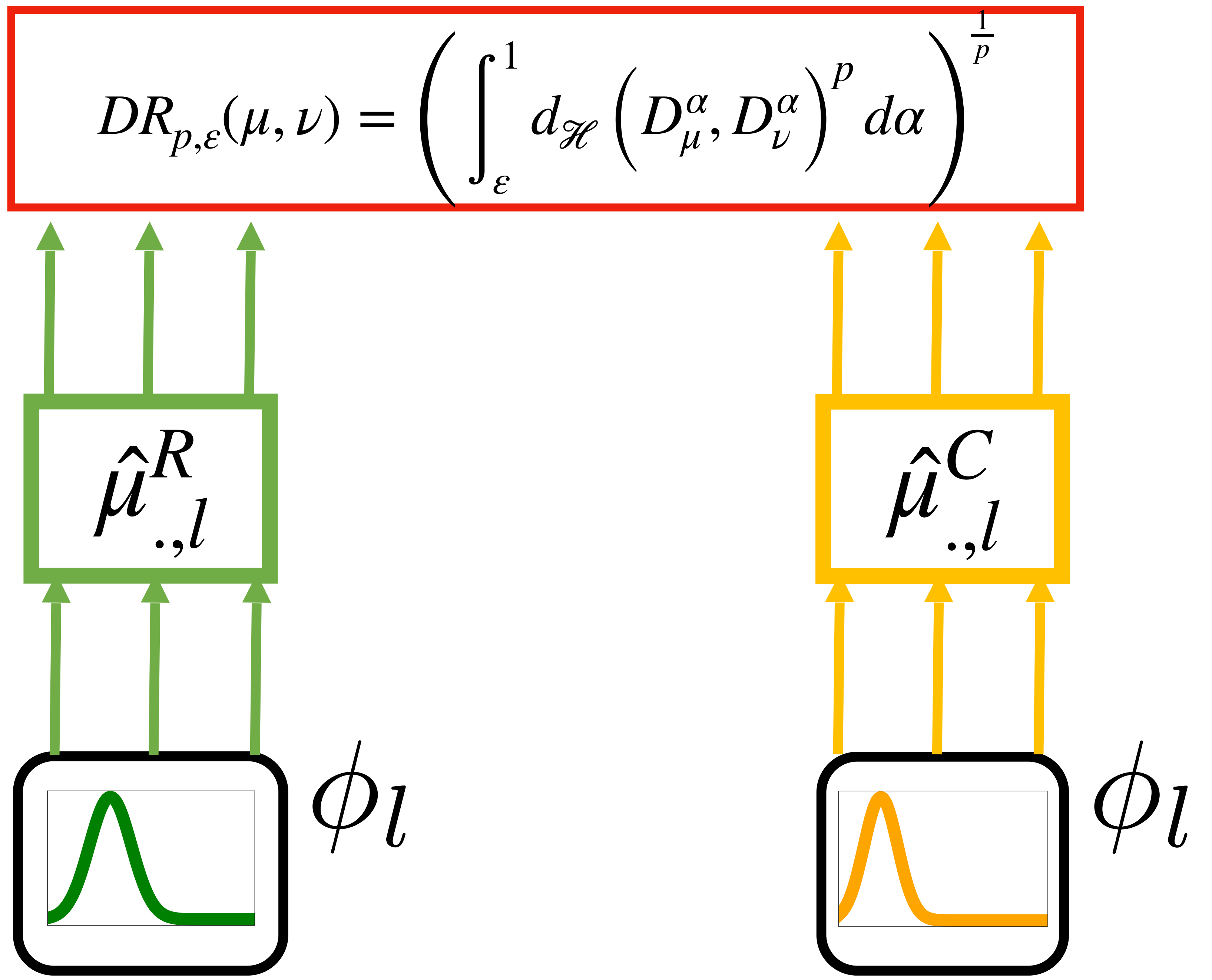
1. Deal with paraphrases
2. Include "semantic"

Limitations

1. Use only one layer

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"

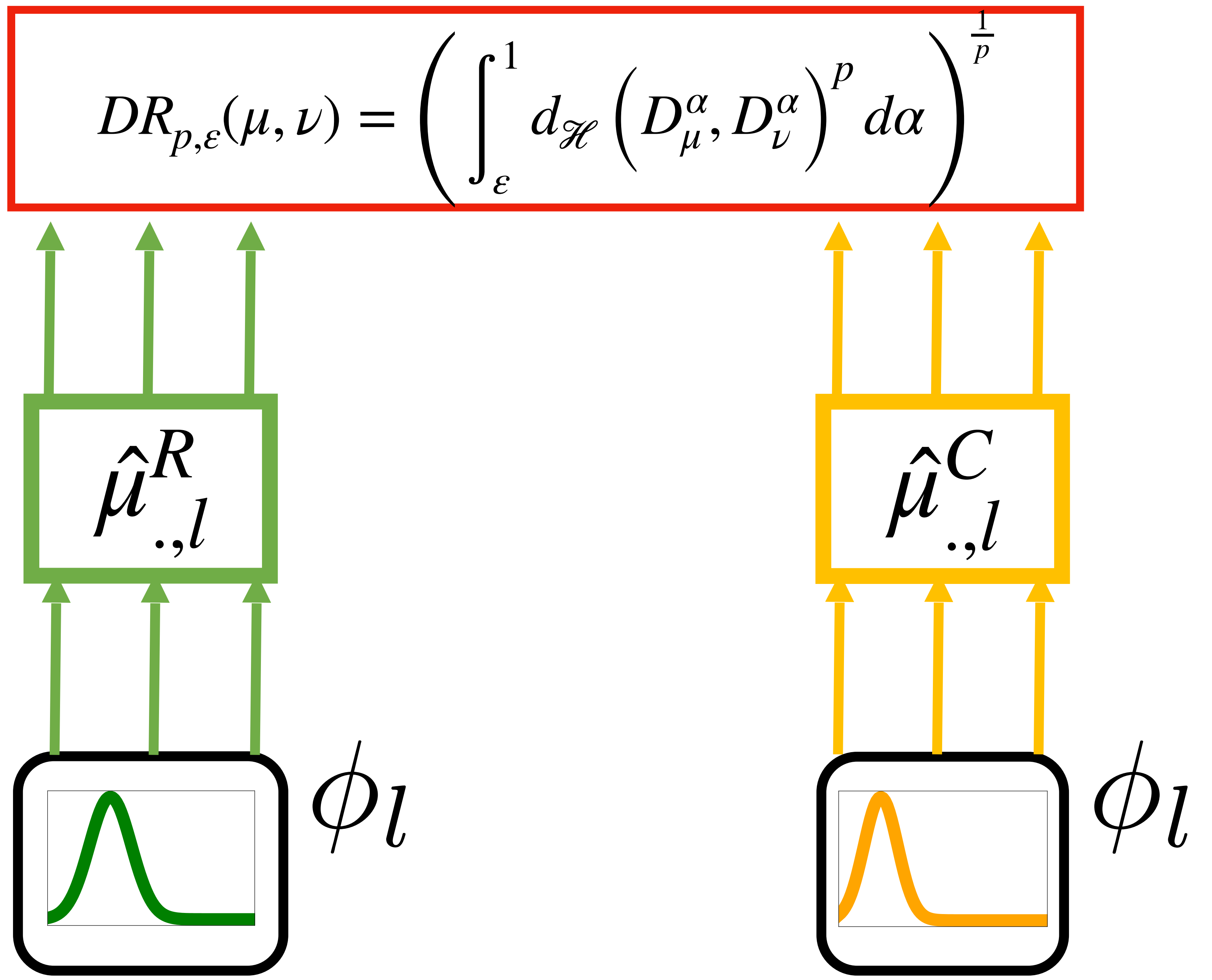
Limitations

1. Use only one layer

Still not interpretable

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

Still not interpretable

Embedding Based Metric With Neural Networks

Embedding Based Metric With Neural Networks

Intuition

R: The weather is cold today.

C: It is freezing today

0.8

Embedding Based Metric With Neural Networks

Intuition

R: The weather is cold today.

C: It is freezing today

0.8

1. Choose your multi-layer encoder

Embedding Based Metric With Neural Networks

Intuition

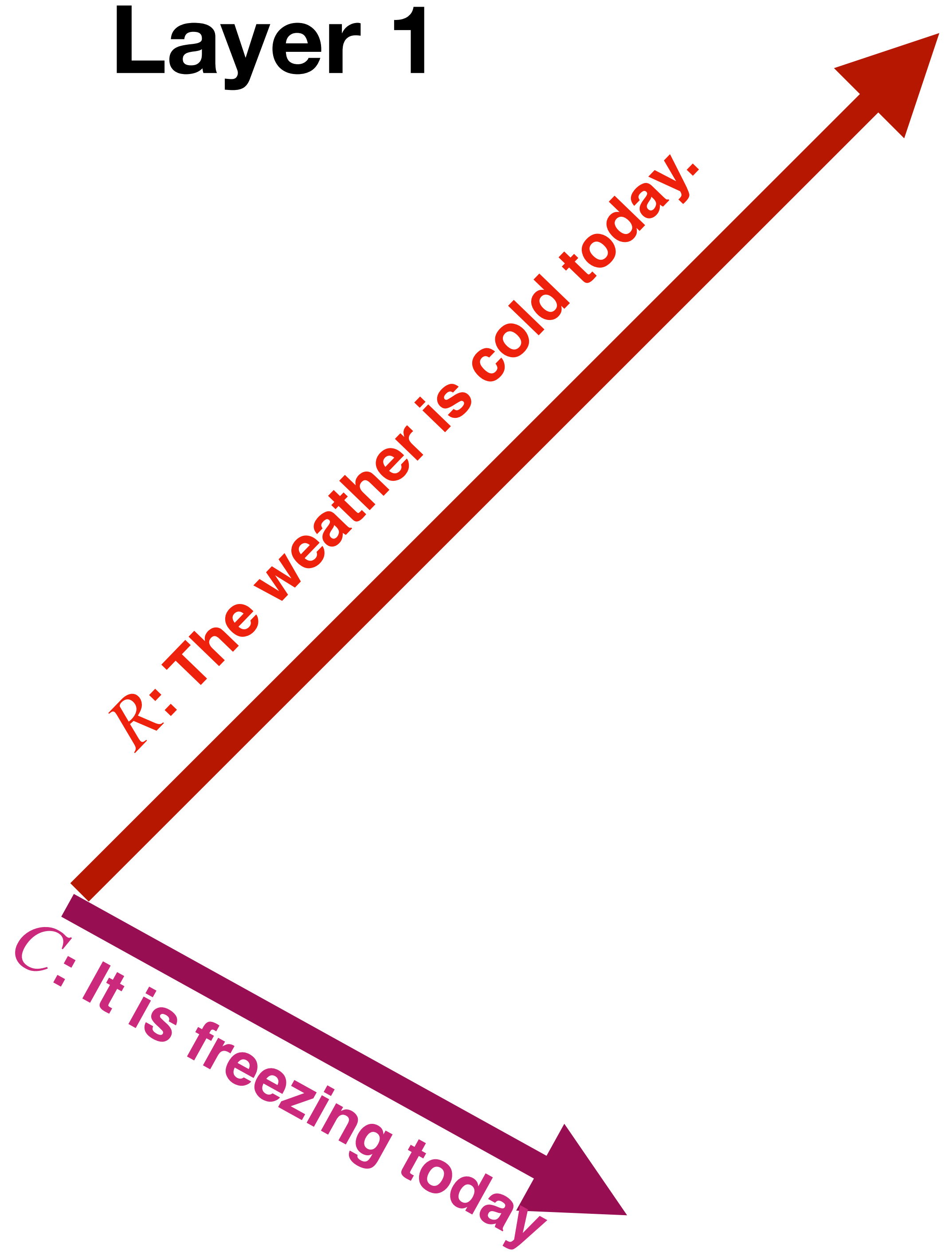
R: The weather is cold today.

C: It is freezing today

0.8

1. Choose your multi-layer encoder

Layer 1



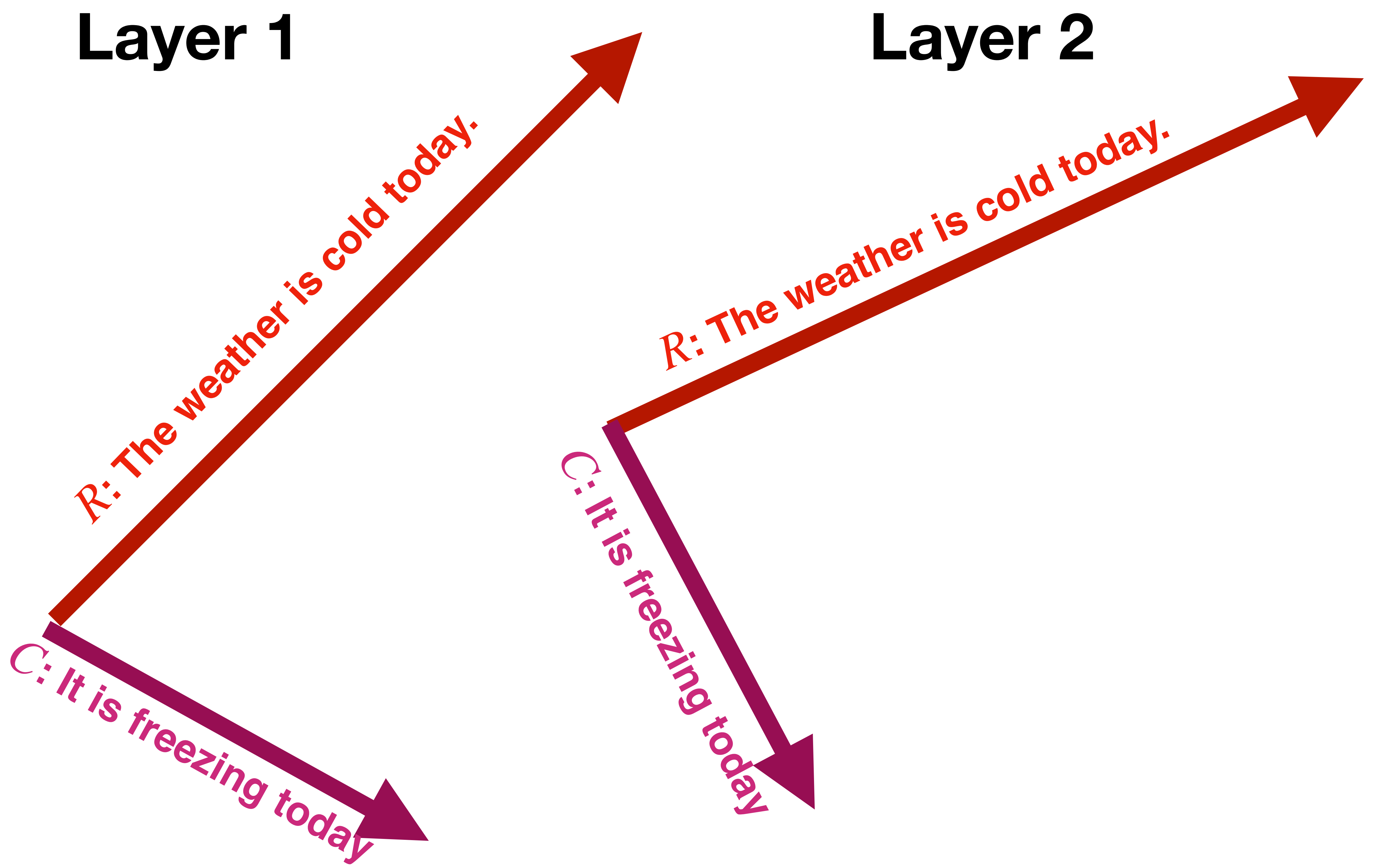
Embedding Based Metric With Neural Networks

Intuition

R: The weather is cold today.

C: It is freezing today

1. Choose your multi-layer encoder



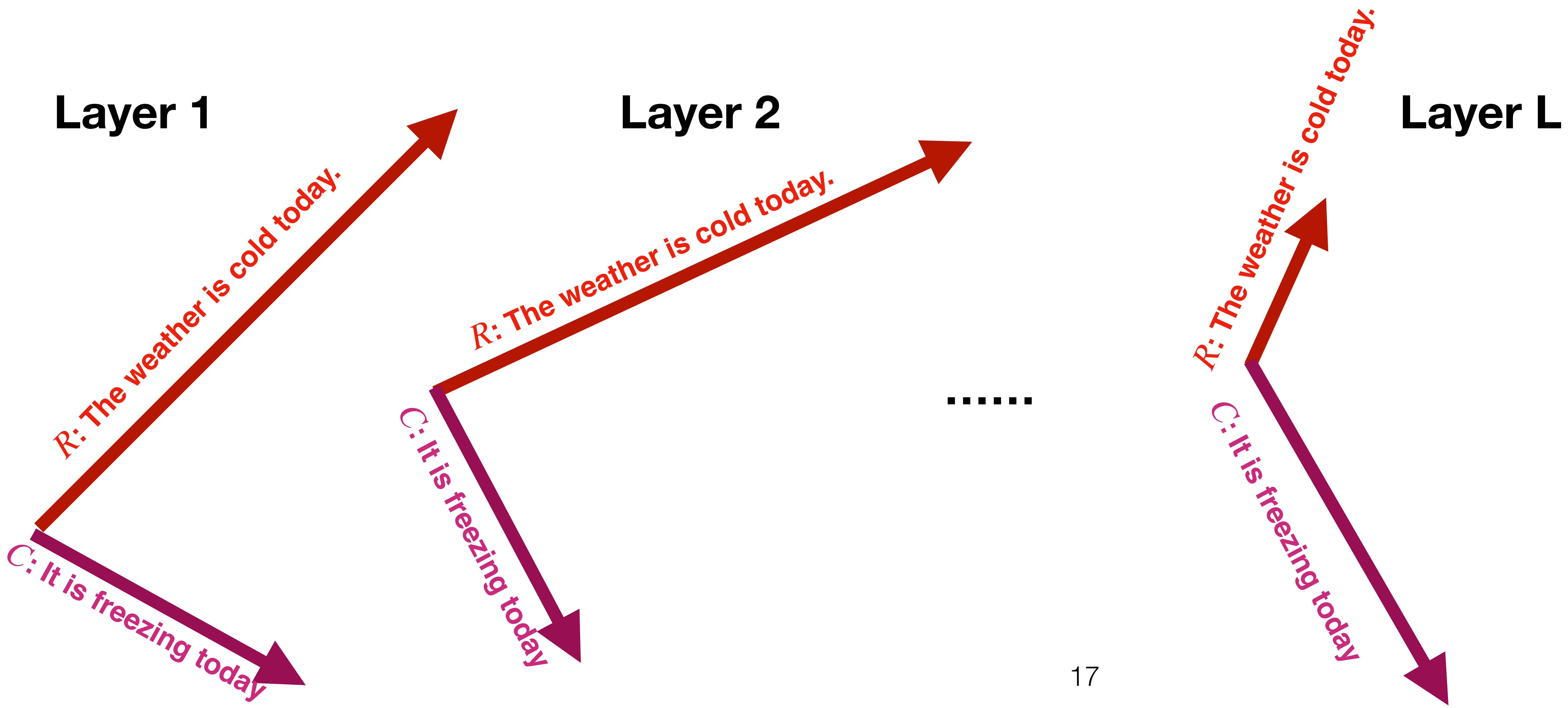
Embedding Based Metric With Neural Networks

Intuition

R: The weather is cold today.

C: It is freezing today

1. Choose your multi-layer encoder



Embedding Based Metric With Neural Networks

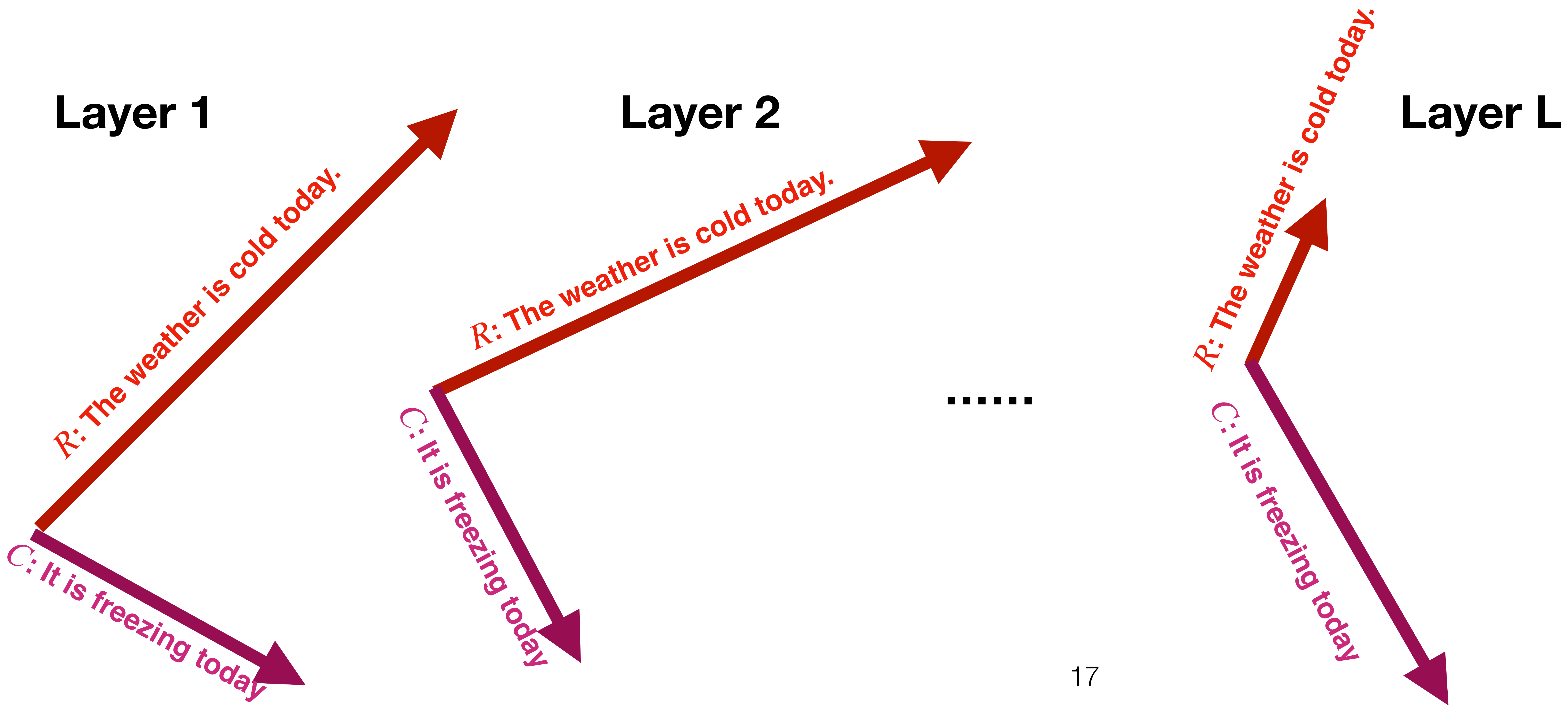
Intuition

R: The weather is cold today.

C: It is freezing today

1. Choose your multi-layer encoder

2. Choose a similarity function euh??



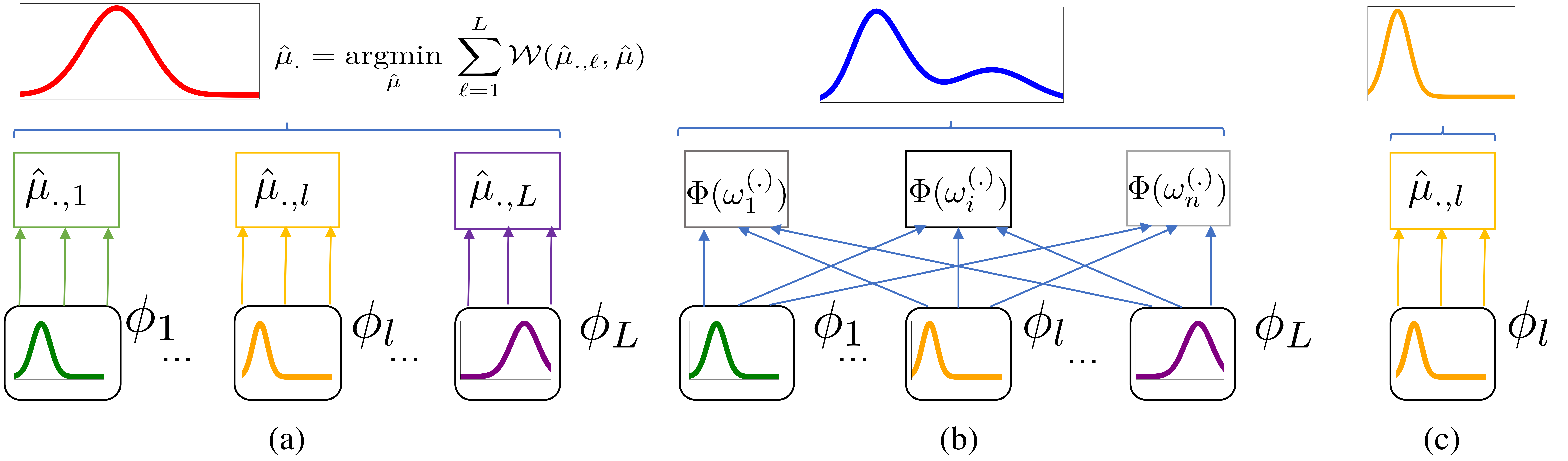
?

BaryScore vs BertScore vs MoverScore

Pierre Colombo, Guillaume Staerman, Chloé Clavel, Pablo Piantanida. Automatic Text Evaluation through the Lens of Wasserstein Barycenters.

BaryScore vs BertScore vs MoverScore

Pierre Colombo, Guillaume Staerman, Chloé Clavel, Pablo Piantanida. Automatic Text Evaluation through the Lens of Wasserstein Barycenters.



BaryScore **MoverScore** **BertScore**

1. How to evaluate Natural Language Generation?

1.1 Context: problems, evaluation of automatic evaluation.

1.2 What are the main metrics to do reference based evaluation of NLG?

1.3 Reference based evaluation of NLG using embedding based metrics.

1.4 Beyond embedding based metrics.

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

Statistical Measures of Similarity

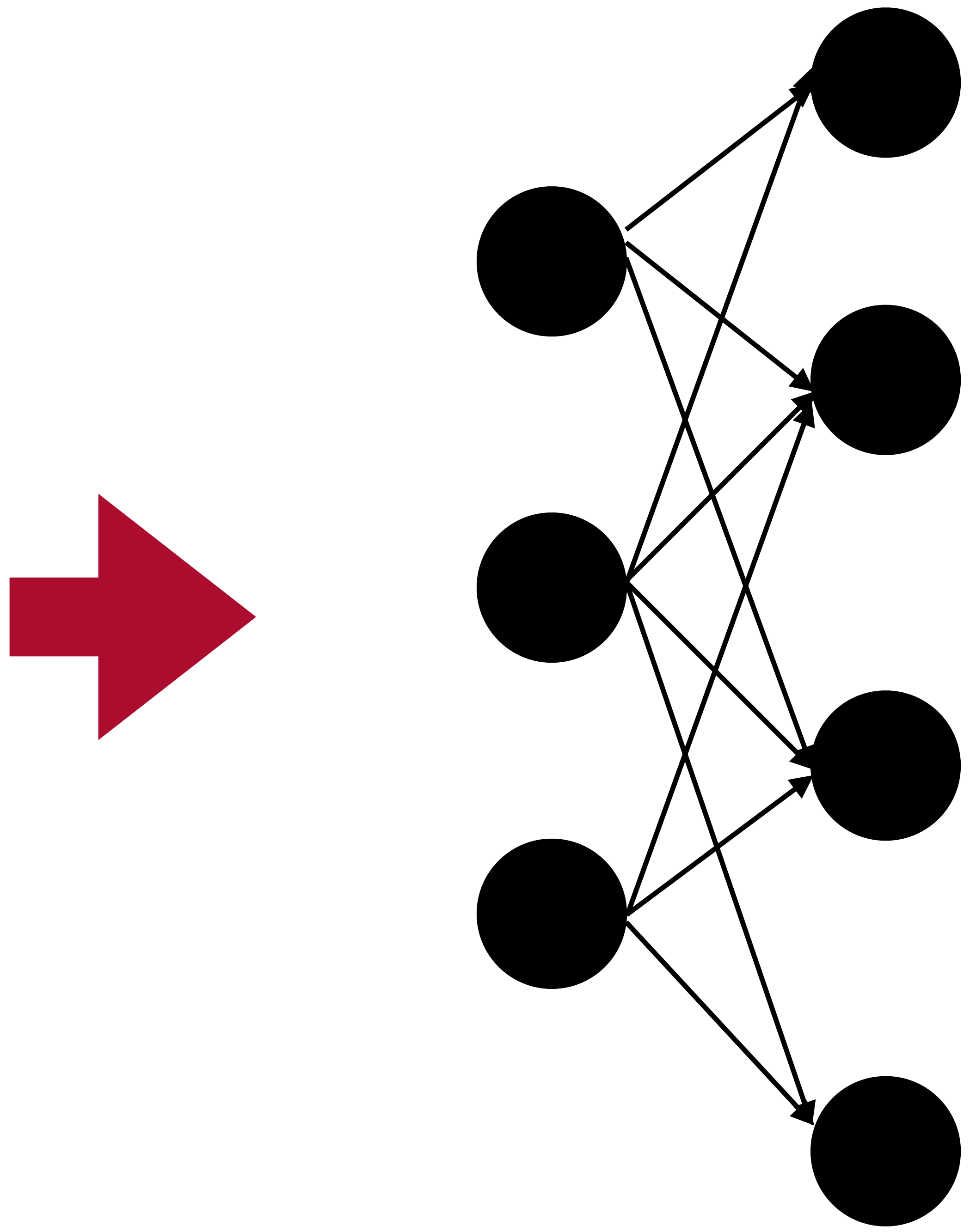
Statistical Measures of Similarity

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

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!



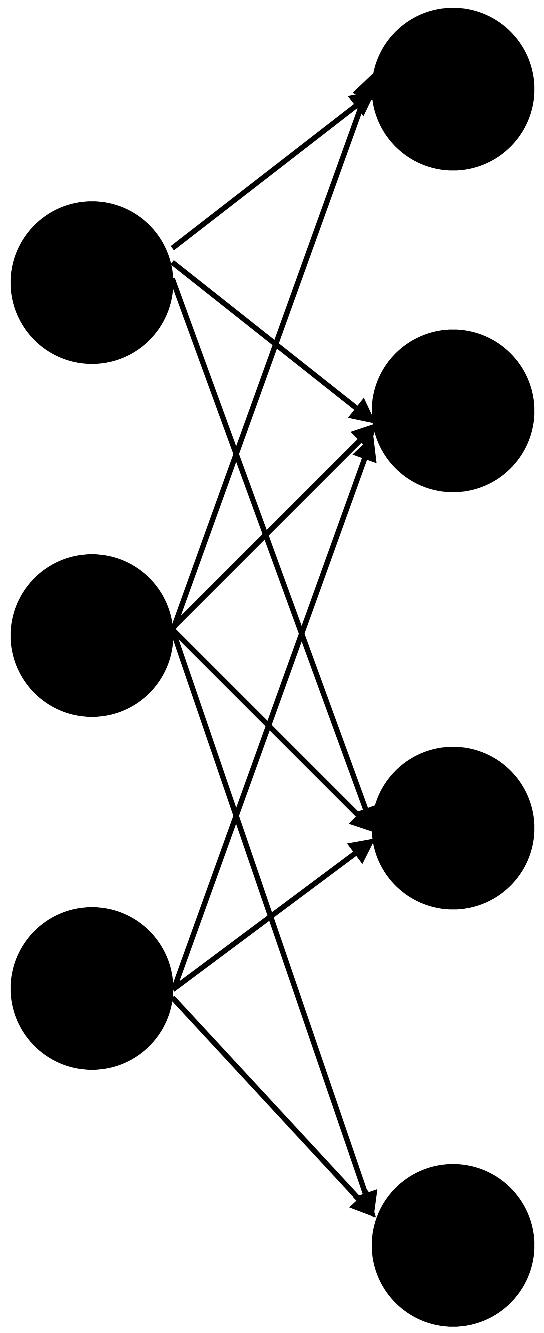
Input Text

Neural Network

Statistical Measures of Similarity

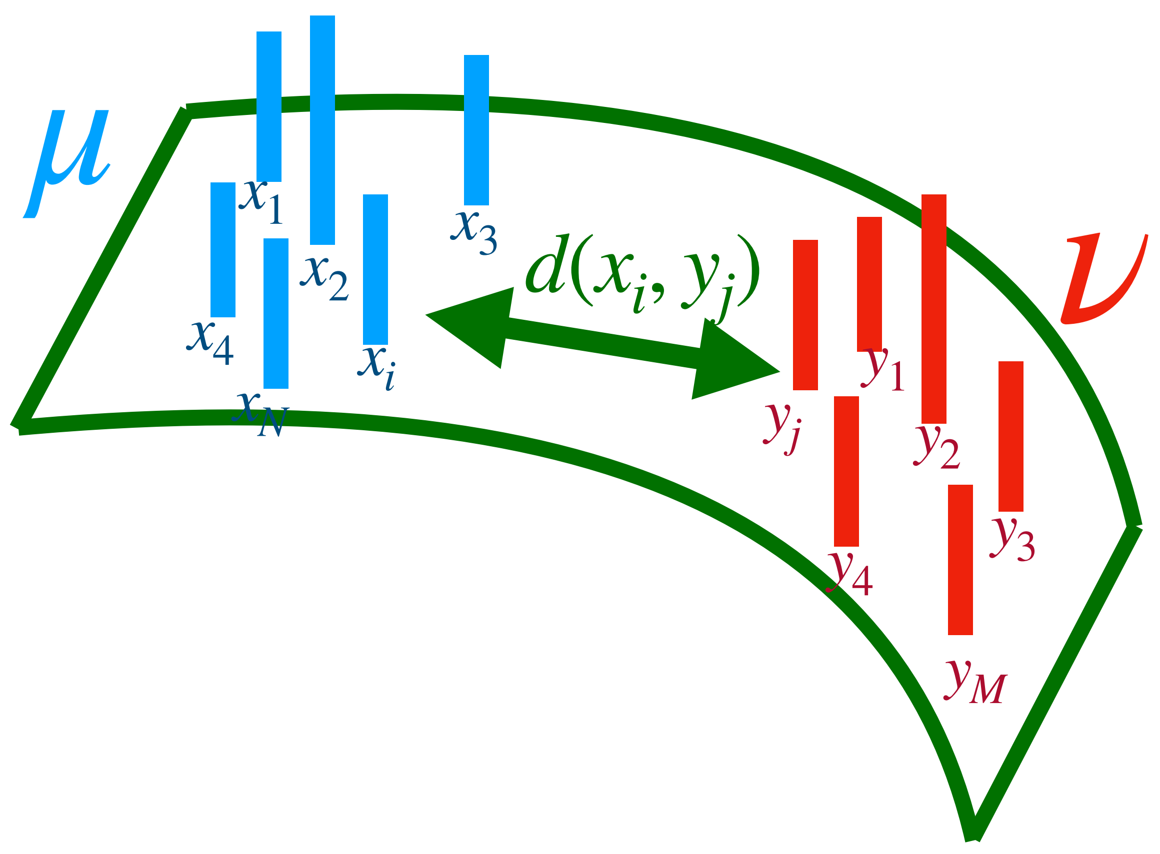
Hello, Chicago.
If there is anyone out
there who still doubts
that America is a place
where all things are
possible, who still
wonders if the dream of
our founders is alive in
our time, [....].
Yes we can!

Input Text



Neural Network

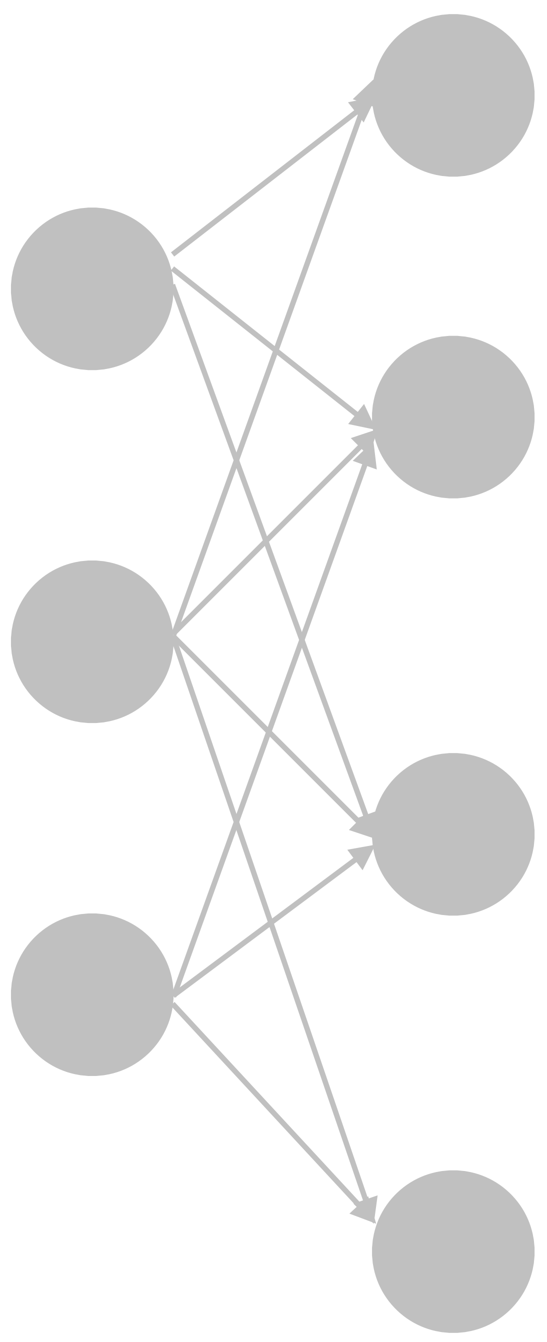
High dimensional data



Statistical Measures of Similarity

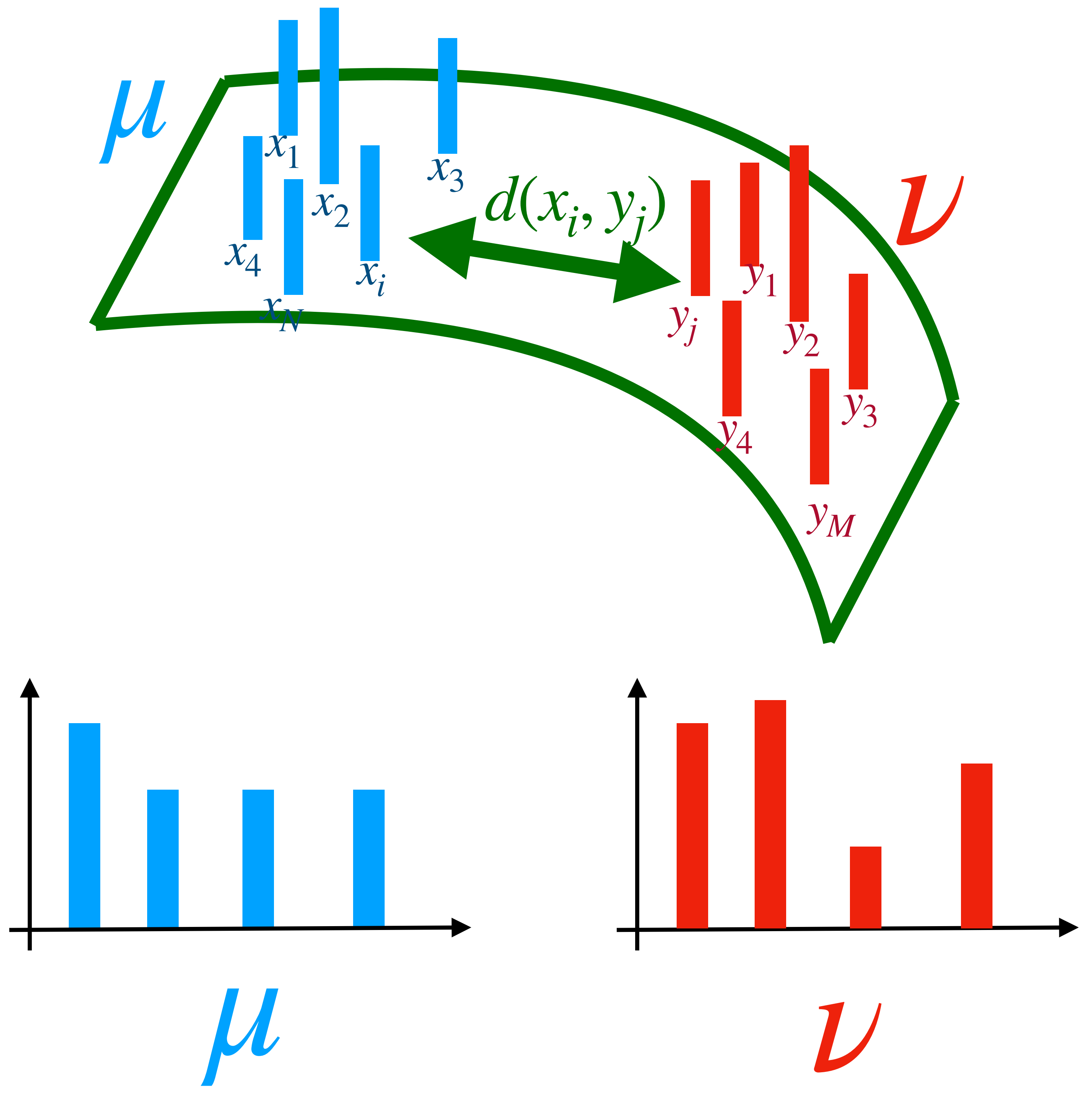
Hello, Chicago.
If there is anyone out there who still doubts that America is a place where all things are possible, who still wonders if the dream of our founders is alive in our time, [….].
Yes we can!

Input Text



Neural Network

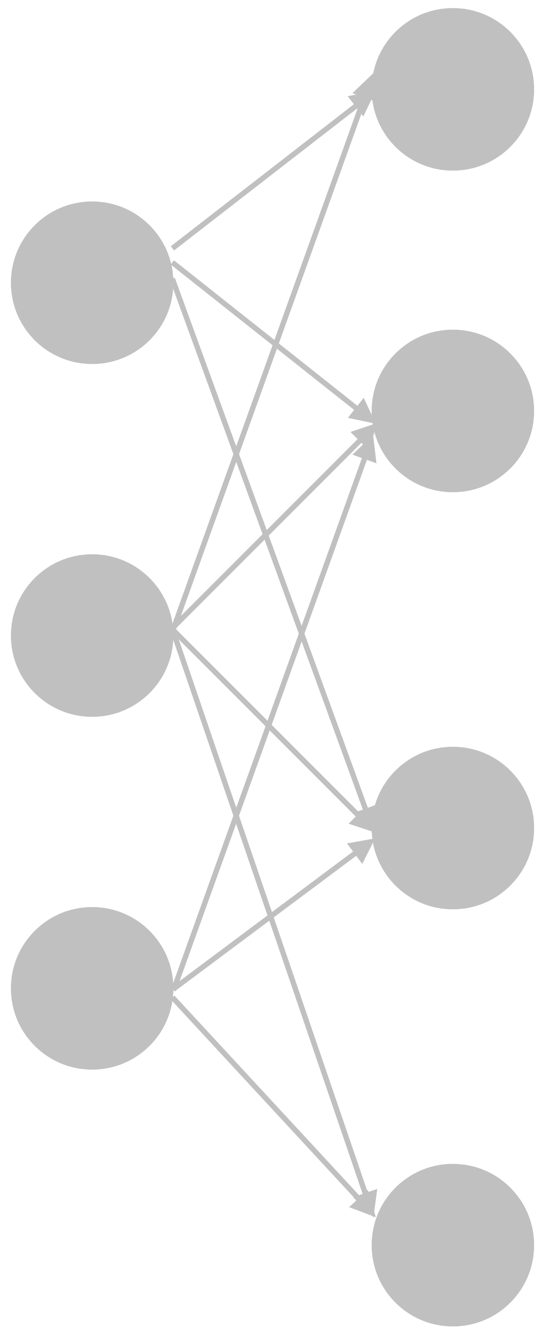
High dimensional data



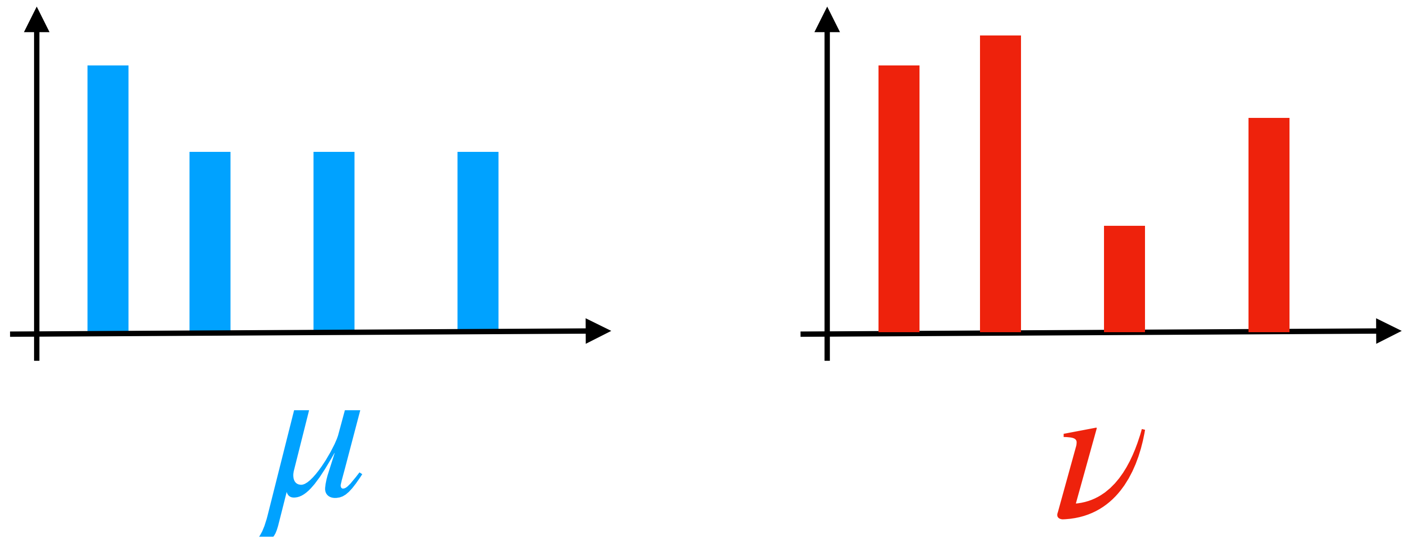
Soft Probabilities

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!



High dimensional data



Input Text

Neural Network

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 !

Existing Methods

InfoLM

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

Assumptions for InfoLM

Assumptions for InfoLM

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

Assumptions for InfoLM

Goal Compute a similarity score between R and C.

Tools Use a pretrained MLM

Assumptions for InfoLM

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

MLM predicts a distribution over Ω

Tools **Use a pretrained MLM**

Assumptions for InfoLM

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

Tools **Use a pretrained MLM**

MLM predicts a distribution over Ω

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

Assumptions for InfoLM

Goal Compute a similarity score between R and C.

Tools Use a pretrained MLM

MLM predicts a distribution over Ω

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

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

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

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

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

Intuition of InfoLM

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

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

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

MLM

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

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

MLM

Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !

Intuition of InfoLM

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

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

MLM predicts a distribution over Ω

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

Similar context

MLM

R: It is **[MASK]** today. $\xrightarrow{p_\Omega(\cdot \,|\, [R]^2)}$

C: It is **[MASK]** this morning ! $\xrightarrow{p_\Omega(\cdot \,|\, [C]^2)}$

Intuition of InfoLM

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

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

MLM predicts a distribution over Ω

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

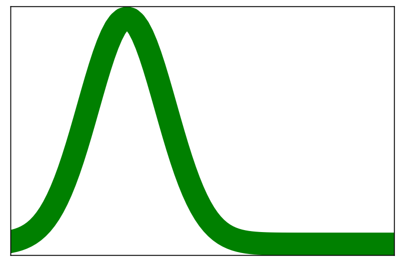
Similar context

R: It is [MASK] today.

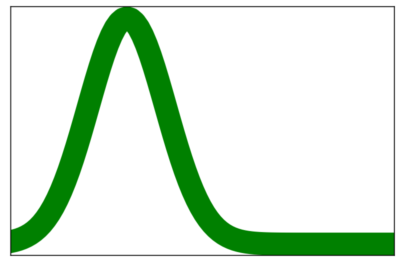
C: It is [MASK] this morning !

MLM

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



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



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

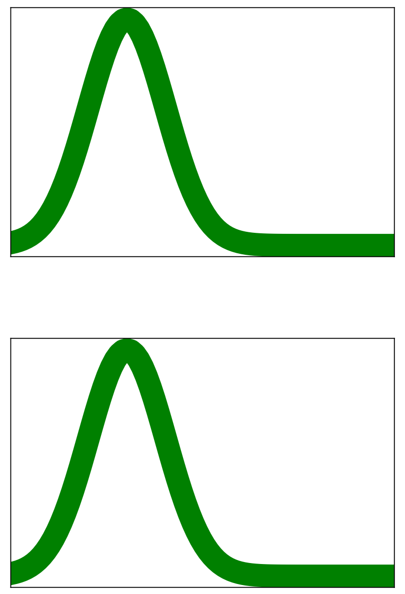
R: It is [MASK] today.

C: It is [MASK] this morning !

MLM

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

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



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

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

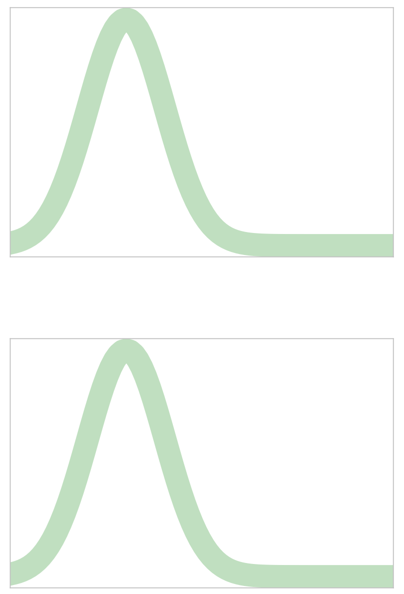
$p_\Omega(\cdot \,|\, [R]^i)$

Similar context

MLM

R: It is [MASK] today. $\xrightarrow{p_\Omega(\cdot \,|\, [R]^2)}$

C: It is [MASK] this morning ! $\xrightarrow{p_\Omega(\cdot \,|\, [C]^2)}$



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

Dissimilar context

R: It is cold [MASK]

C: It is [MASK] this morning !

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

MLM

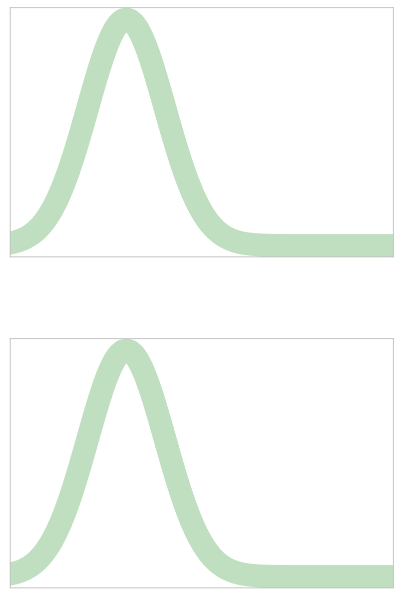
Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !

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

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



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

Dissimilar context

R: It is cold [MASK]

C: It is [MASK] this morning !

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

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

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

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

MLM

Similar context

R: It is [MASK] today.

C: It is [MASK] this morning !

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

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

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

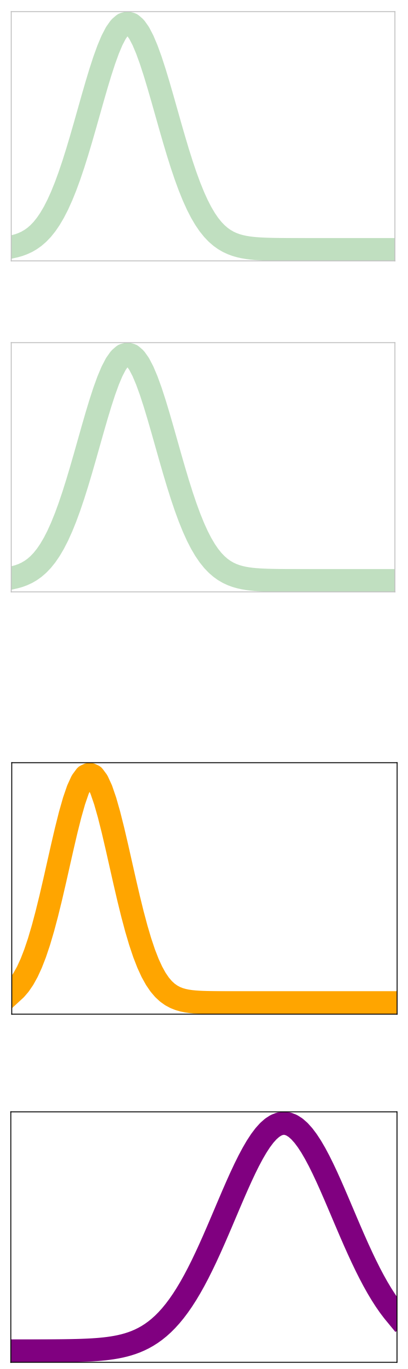
Dissimilar context

R: It is cold [MASK]

C: It is [MASK] this morning !

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

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



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

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

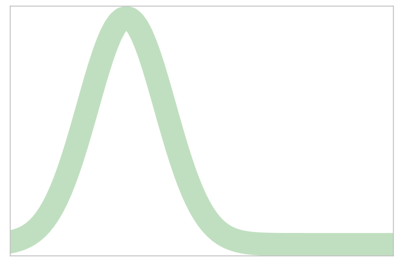
MLM

Similar context

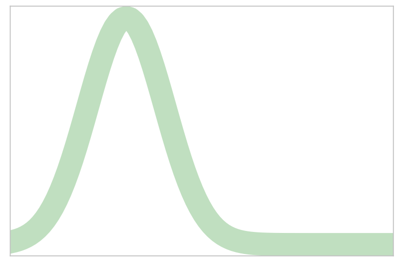
R: It is [MASK] today.

C: It is [MASK] this morning !

$p_\Omega(\cdot \mid [R]^2) \rightarrow$



$p_\Omega(\cdot \mid [C]^2) \rightarrow$



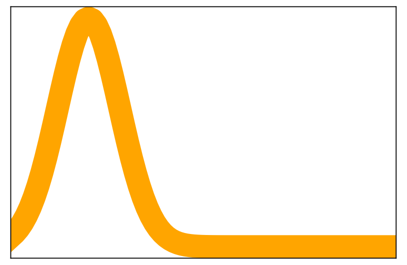
$\mathcal{I}\left(p_\Omega(\cdot \mid [R]^2), p_\Omega(\cdot \mid [C]^2)\right) \sim 0$

Dissimilar context

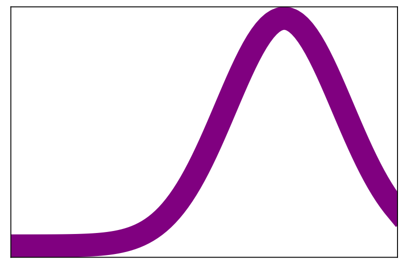
R: It is cold [MASK]

C: It is [MASK] this morning !

$p_\Omega(\cdot \mid [R]^3) \rightarrow$



$p_\Omega(\cdot \mid [C]^2) \rightarrow$



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

Context Aggregation

Context Aggregation

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

Context Aggregation

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

How to aggregate contexts?

Context Aggregation

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

How to aggregate contexts? →

Context Aggregation

Goal Compute a similarity score between R and C.

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

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]

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]

Candidate

[MASK] is freezing this morning !

...

It is [MASK] this morning !

...

It is freezing this morning [MASK]

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]

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

Candidate

[MASK] is freezing this morning !

...

It is [MASK] this morning !

...

It is freezing this morning [MASK]

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]

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

Candidate

[MASK] is freezing this morning !

...

It is [MASK] this morning !

...

It is freezing this morning [MASK]

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

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

Experimental Setting

Data2text Generation

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

Gardent et al. 2017

Ferreira et al. (2020)

Perez-Beltrachini et al 2016

Experimental Setting

Data2text Generation

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

Gardent et al. 2017

Ferreira et al. (2020)

Perez-Beltrachini et al 2016

Summary Generation

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

Nallapati et al. 2016)

Bhandari et al. (2020)

Nenkova and Passonneau 2004

Experimental Setting

Data2text Generation

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

Gardent et al. 2017

Ferreira et al. (2020)

Perez-Beltrachini et al 2016

Summary Generation

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

Nallapati et al. 2016)

Bhandari et al. (2020)

Nenkova and Passonneau 2004

Results

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

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

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

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

Takeaways of the first part:

Takeaways of the first part:

We explored different metrics for automatic NLG evaluation

Takeaways of the first part:

We explored different metrics for automatic NLG evaluation

Embedding Based

Soft Probability based

Takeaways of the first part:

We explored different metrics for automatic NLG evaluation

Embedding Based

Soft Probability based

Different Metrics correlate better with different human criterion

Takeaways of the first part:

We explored different metrics for automatic NLG evaluation

Embedding Based

Soft Probability based

Different Metrics correlate better with different human criterion

One task: If we want to have an exhaustive evaluation we need to consider several metrics.

Multitask: To evaluate on system on different tasks we need different metrics (data2text vs Translation)

Takeaways of the first part:

We explored different metrics for automatic NLG evaluation

Embedding Based

Soft Probability based

Different Metrics correlate better with different human criterion

One task: If we want to have an exhaustive evaluation we need to consider several metrics.

Multitask: To evaluate on system on different tasks we need different metrics (data2text vs Translation)

Let's speak about how to aggregate different metrics to obtain stronger evaluation procedures.

2. How to aggregate several metrics?

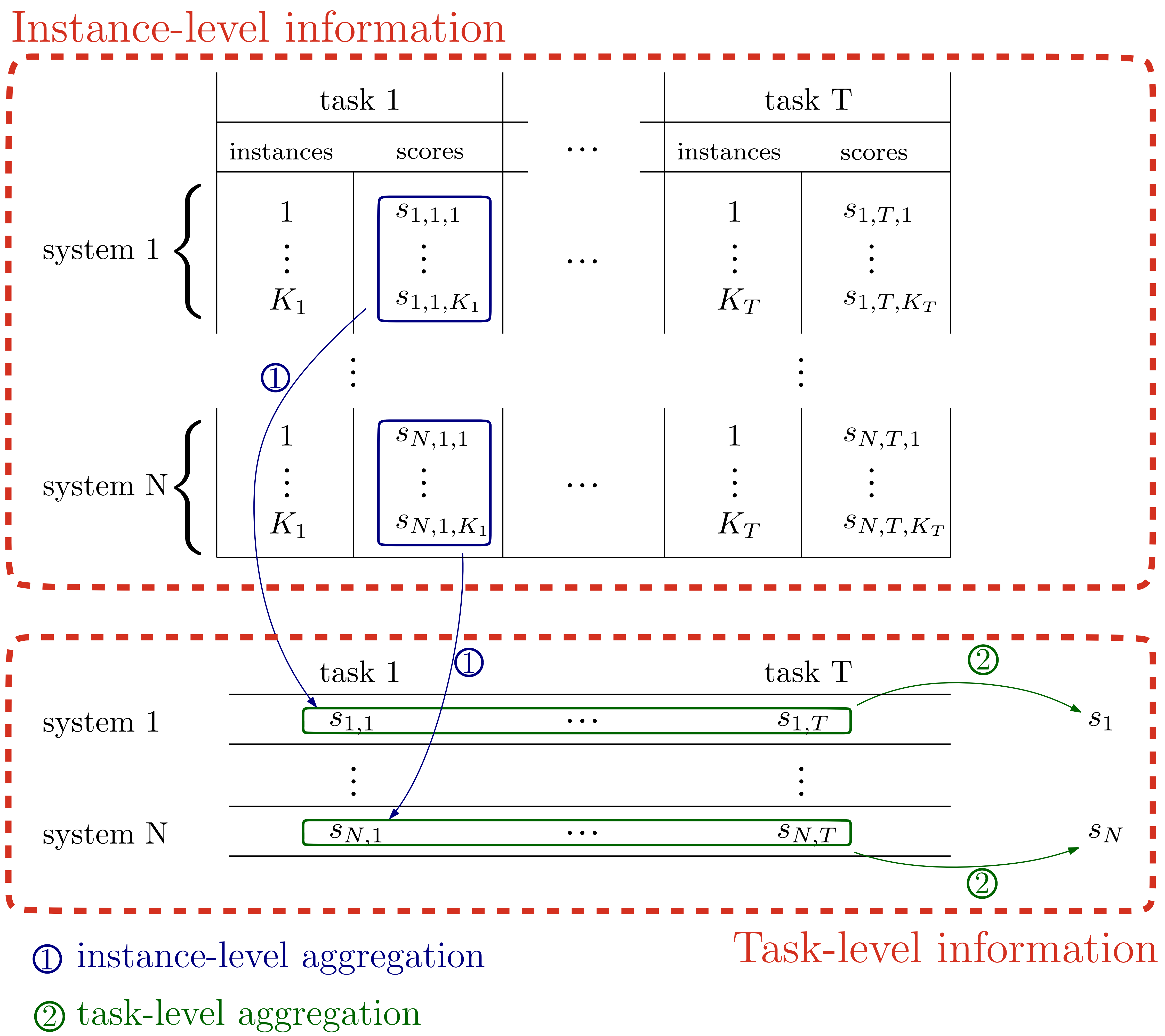
1.1 Framework

1.2 Task Level Aggregation

1.3 Instance Level Aggregation

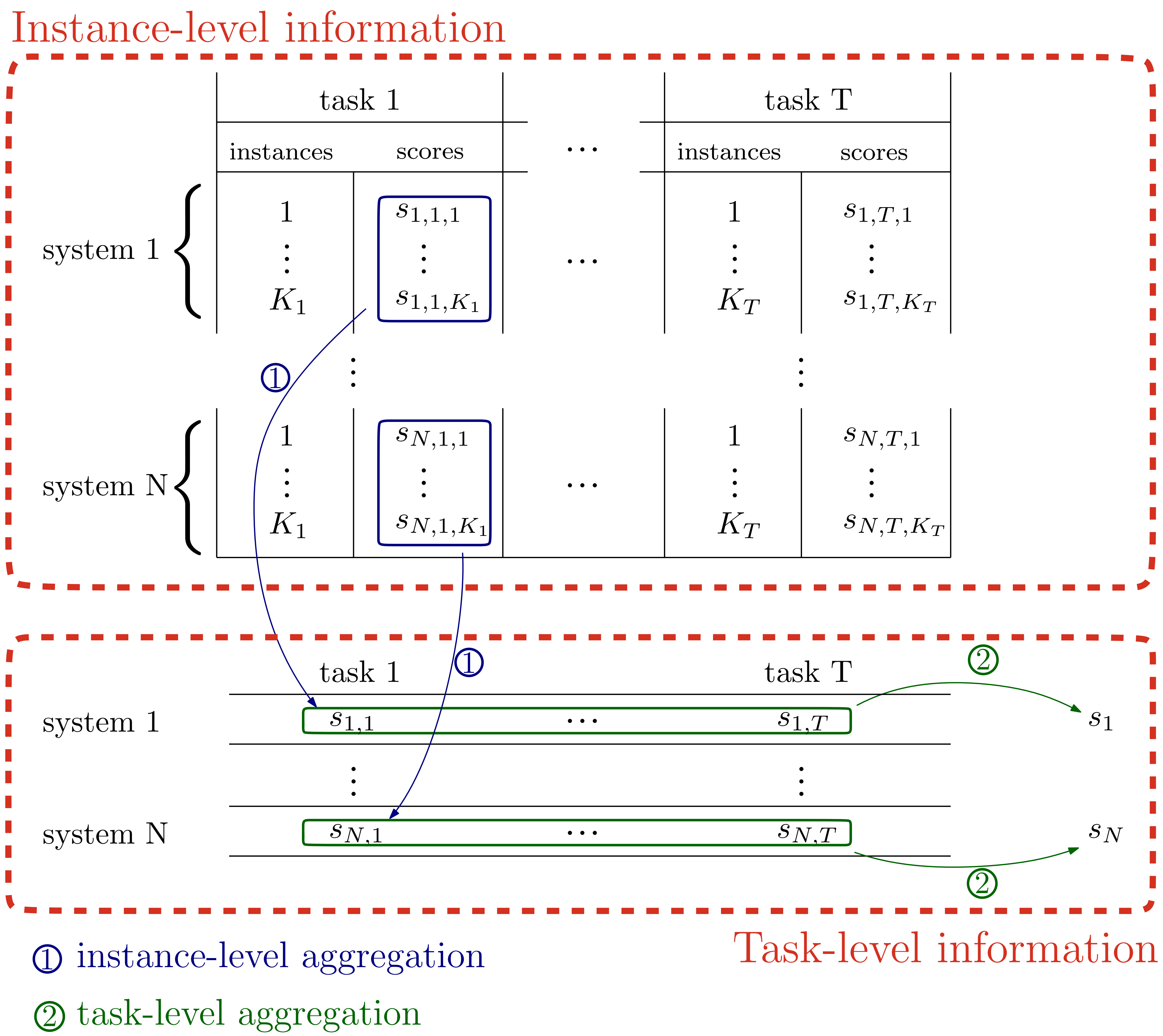
Pierre Colombo, Nathan Noiry, Ekhine Irurozki and Stephan Clemencon. What are the best Systems? New Perspectives on NLP Benchmarking.

Framework



Instance-level information
task 1
task T
instances
scores
...
system 1
1
K_1
$s_{1,1,1}$
$s_{1,1,K_1}$
$s_{1,T,1}$
s_{1,T,K_T}
K_T
system N
$s_{N,1,1}$
$s_{N,1,K_1}$
$s_{N,T,1}$
s_{N,T,K_T}
①
②
$s_{1,1}$
$s_{1,T}$
$s_{N,1}$
$s_{N,T}$
s_1
s_N
Task-level information
① instance-level aggregation
② task-level aggregation

Framework



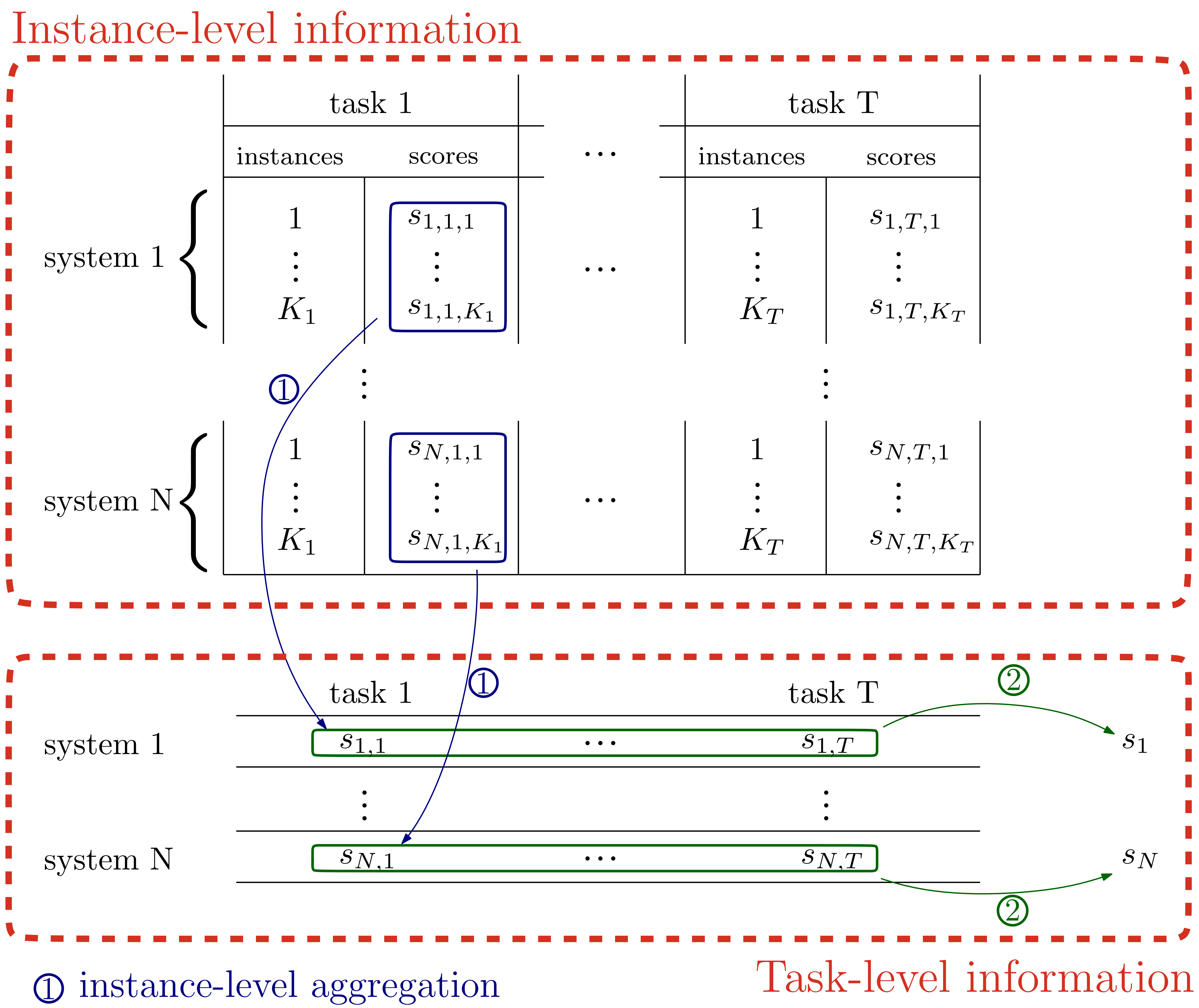
Setting:

1. One has access to the scores of N systems across T tasks.

2. Each task t being associated with a metric and a test set of size K_t.

3. We have $s_{n,t,k} \in \mathbb{R}$

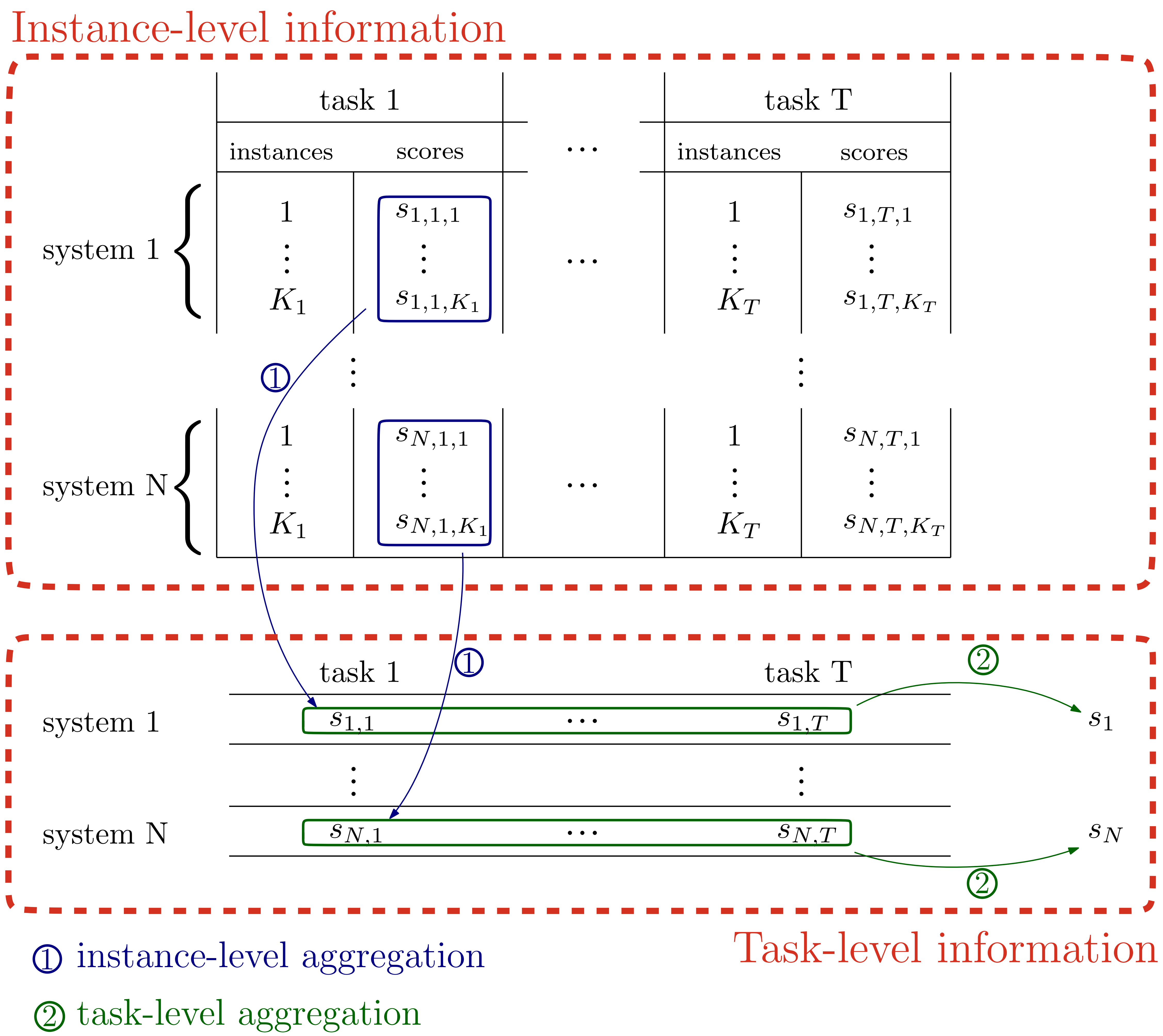
First problem: task-level ranking



① instance-level aggregation

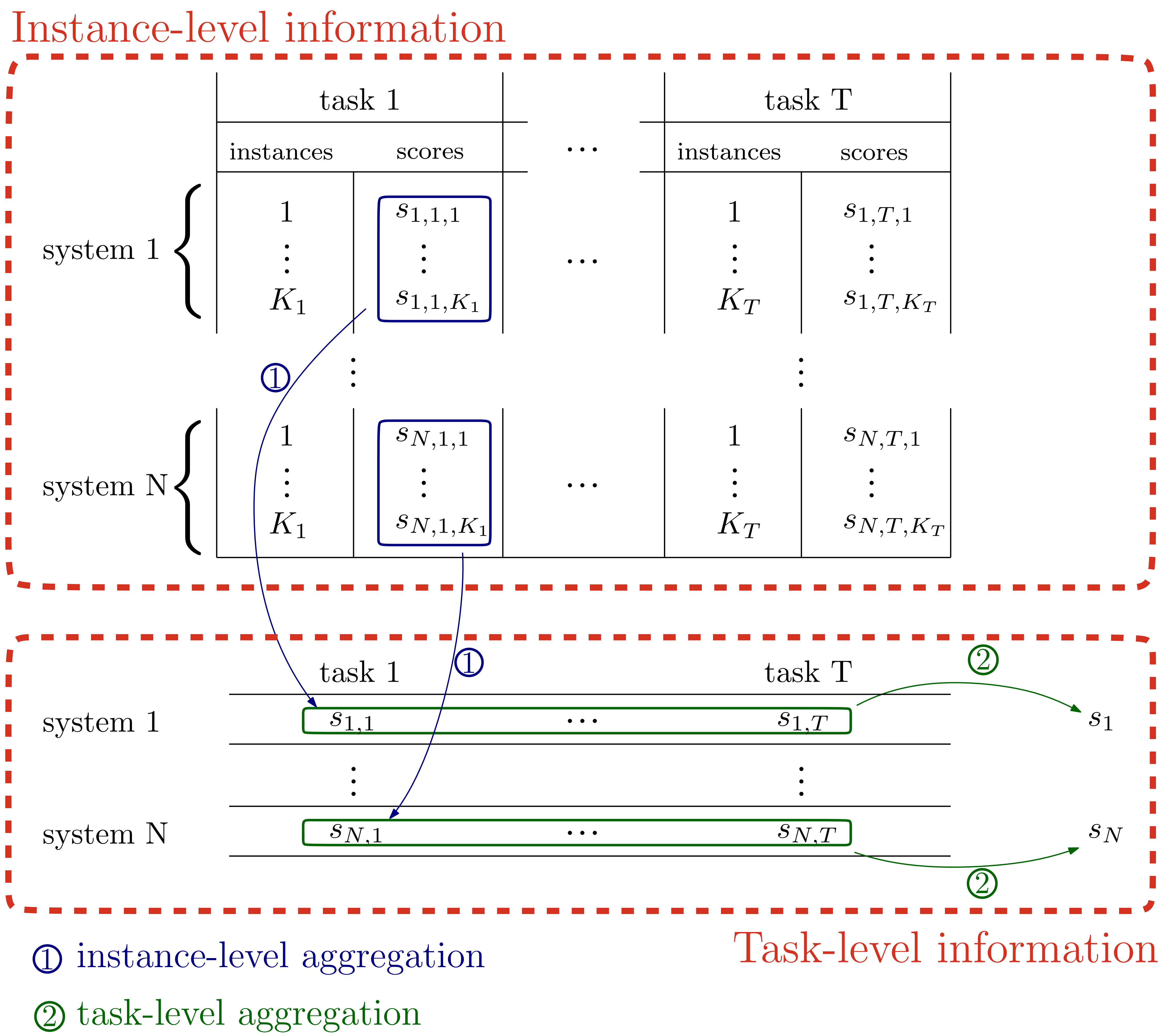
② task-level aggregation

First problem: task-level ranking



For every n and every t, we only have access to the aggregated performance of system n on task t

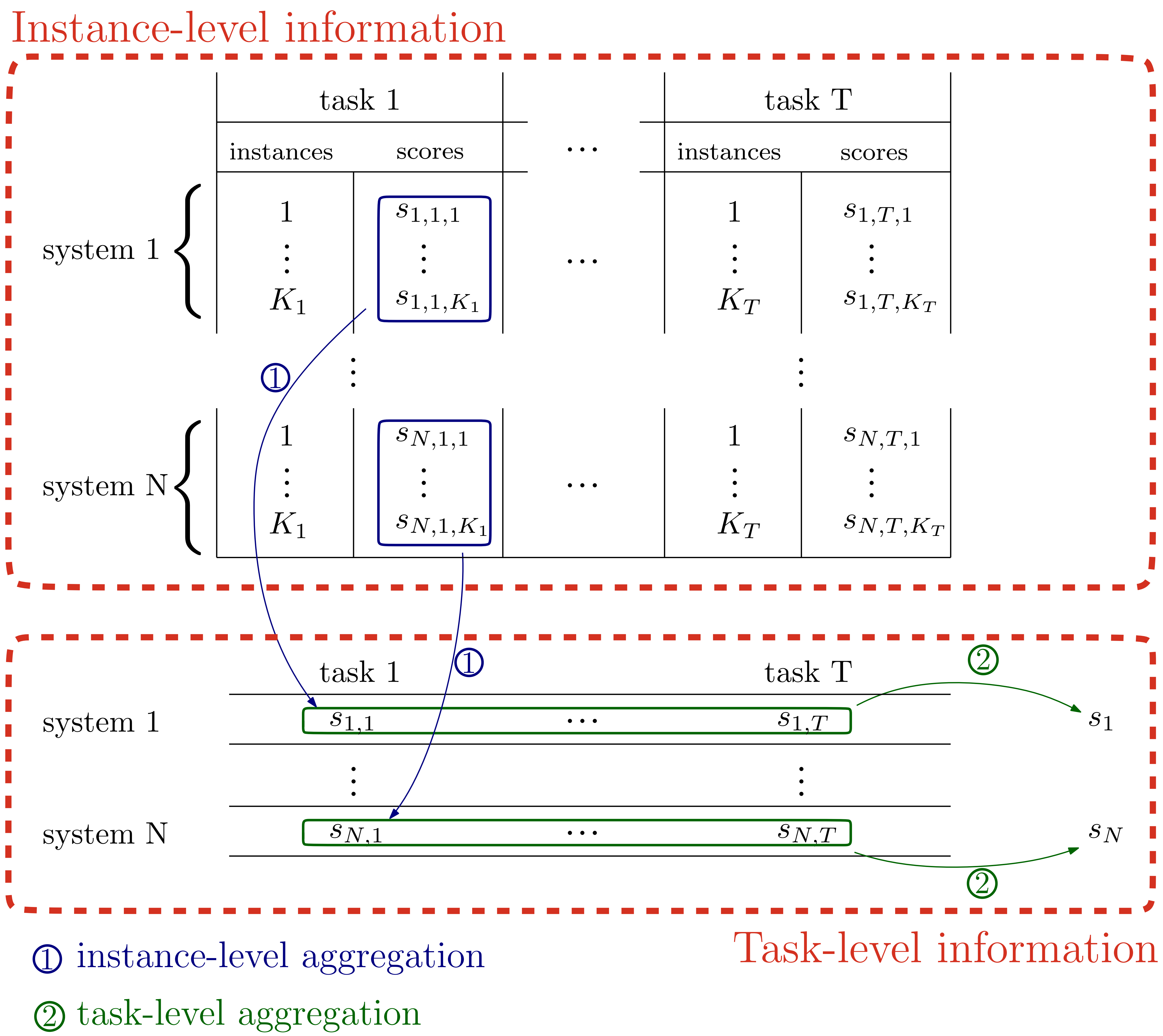
First problem: task-level ranking



For every n and every t, we only have access to the aggregated performance of system n on task t

$$s_{n,t} \in \mathbb{R}$$

First problem: task-level ranking



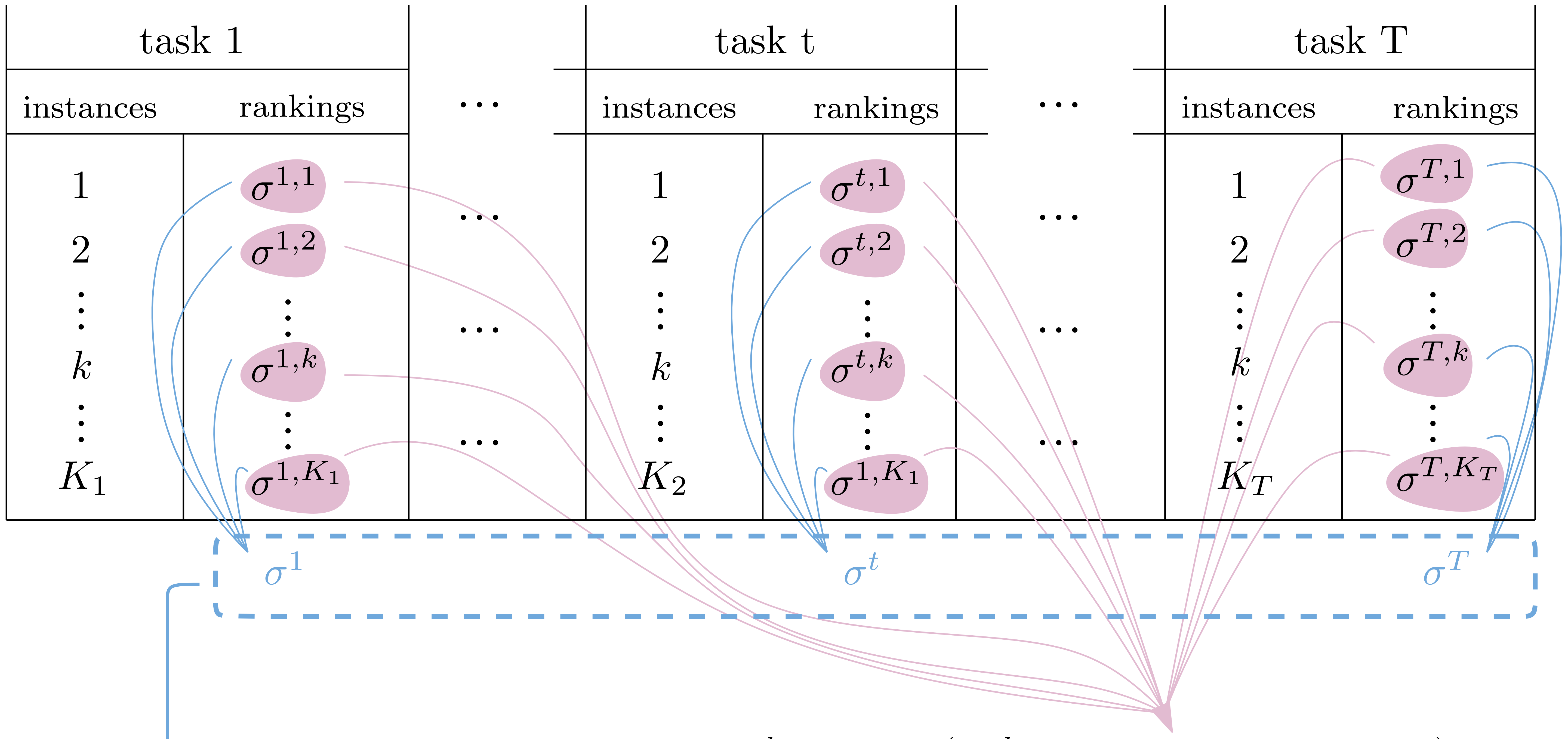
For every n and every t, we only have access to the aggregated performance of system n on task t

$$s_{n,t} \in \mathbb{R}$$

Goal: find an aggregation procedure that orders the systems.

Second problem: instance-level ranking

Second problem: instance-level ranking

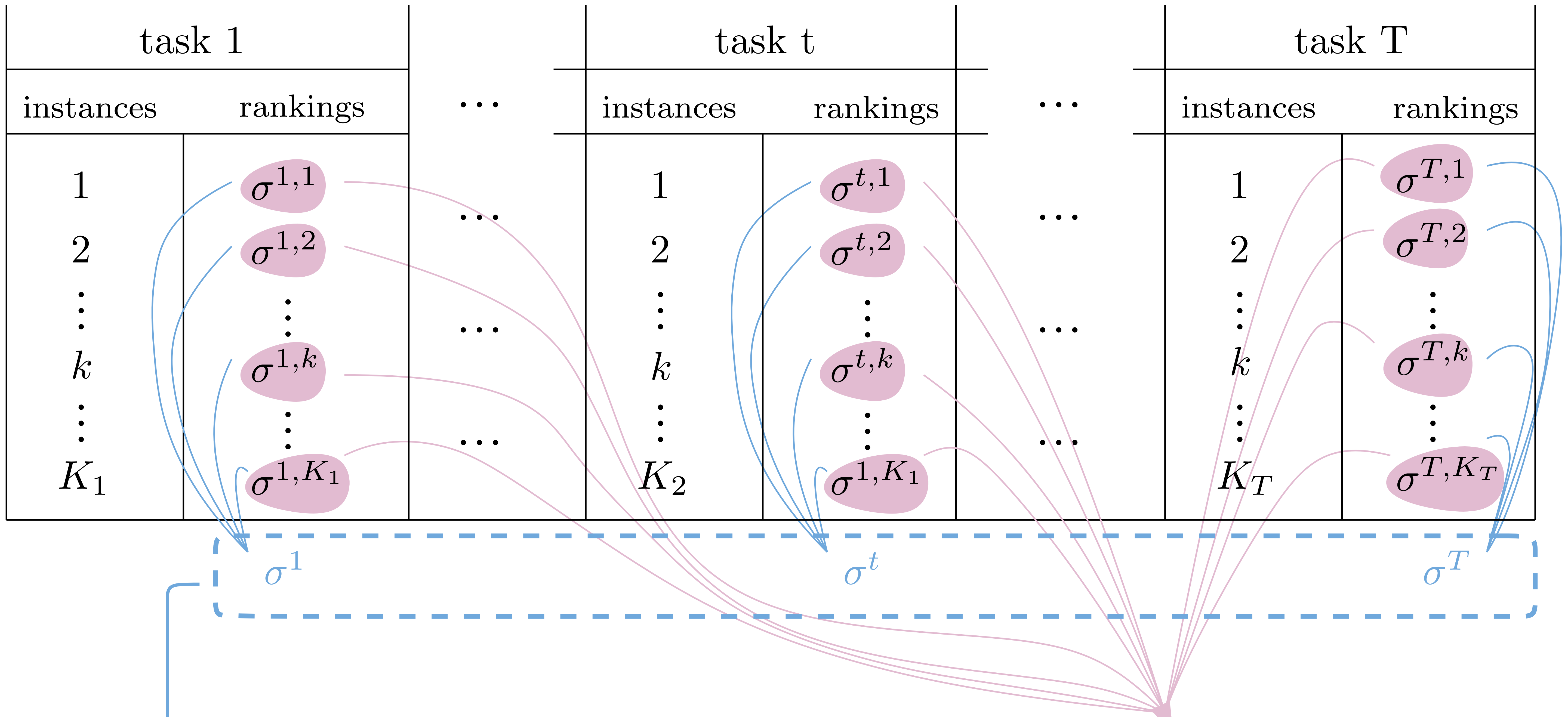


One level aggregation: $\sigma^l = \text{Borda}\big(\sigma^{t,k},\, 1 \le t \le T,\, 1 \le k \le K_t\big)$

Two level aggregation: $\forall 1 \le t \le T, \quad \sigma^t = \text{Borda}\big(\sigma^{t,k},\, 1 \le k \le K_t\big)$

$\sigma^{2l} = \text{Borda}\big(\sigma^t,\, 1 \le t \le T\big)$

Second problem: instance-level ranking



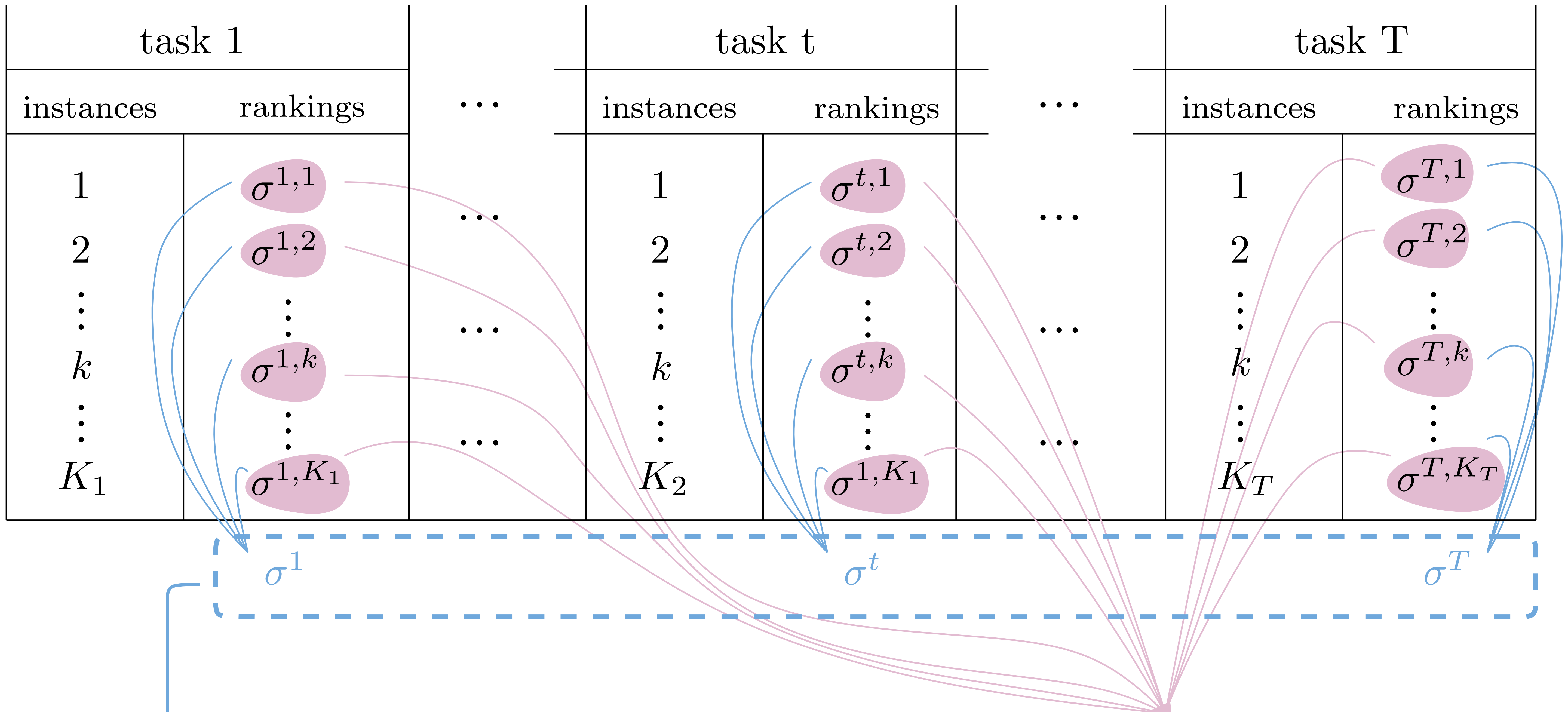
One level aggregation: $\sigma^l = \text{Borda}\big(\sigma^{t,k},\, 1 \leq t \leq T,\, 1 \leq k \leq K_t\big)$

Two level aggregation: $\forall 1 \leq t \leq T, \quad \sigma^t = \text{Borda}\big(\sigma^{t,k},\, 1 \leq k \leq K_t\big)$

$\sigma^{2l} = \text{Borda}\big(\sigma^t,\, 1 \leq t \leq T\big)$

For every n, every t and every k, access to the aggregated performance of system n on instance k of task t

Second problem: instance-level ranking



One level aggregation: $\sigma^l = \text{Borda}\big(\sigma^{t,k},\, 1 \leq t \leq T,\, 1 \leq k \leq K_t\big)$

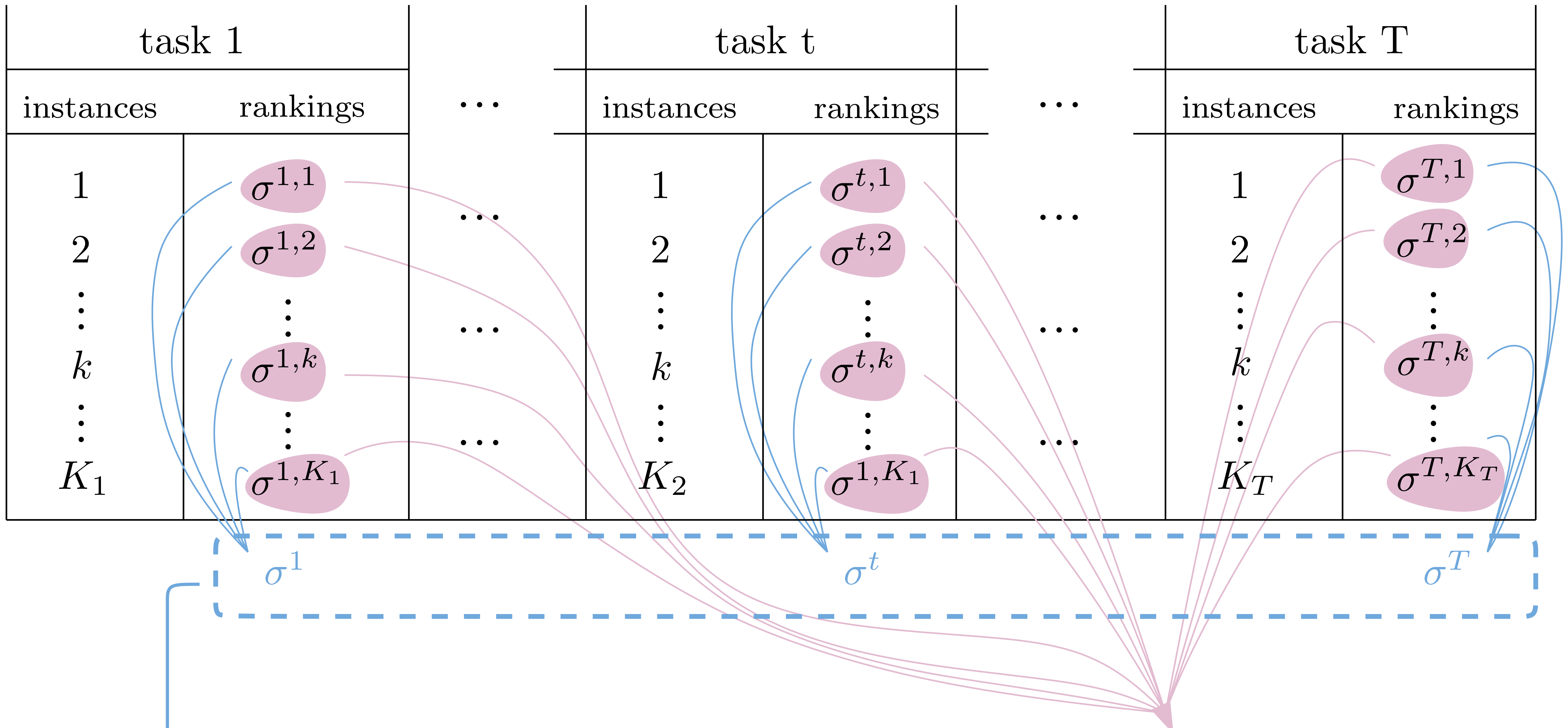
Two level aggregation: $\forall 1 \leq t \leq T, \quad \sigma^t = \text{Borda}\big(\sigma^{t,k},\, 1 \leq k \leq K_t\big)$

$\sigma^{2l} = \text{Borda}\big(\sigma^t,\, 1 \leq t \leq T\big)$

For every n, every t and every k, access to the aggregated performance of system n on instance k of task t

$$s_{n,t,k} \in \mathbb{R}$$

Second problem: instance-level ranking



For every n, every t and every k, access to the aggregated performance of system n on instance k of task t

$$s_{n,t,k} \in \mathbb{R}$$

One level aggregation: $\sigma^l = \mathrm{Borda}\big(\sigma^{t,k},\ 1 \leq t \leq T,\ 1 \leq k \leq K_t\big)$

Two level aggregation: $\forall 1 \leq t \leq T, \quad \sigma^t = \mathrm{Borda}\big(\sigma^{t,k},\ 1 \leq k \leq K_t\big)$

$\sigma^{2l} = \mathrm{Borda}\big(\sigma^t,\ 1 \leq t \leq T\big)$

Goal: find an aggregation procedure that orders the systems.

2. How to aggregate several metrics?

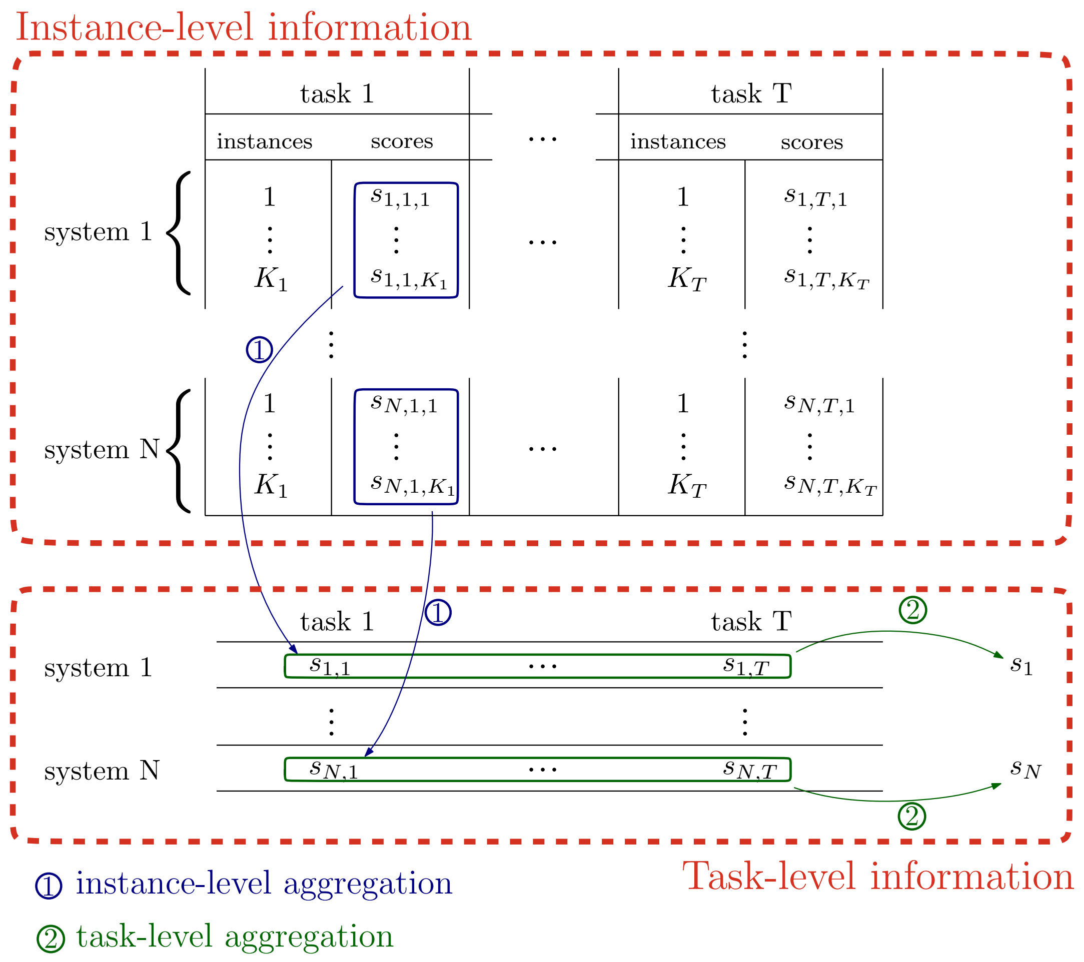
1.1 Framework

1.2 Task Level Aggregation

1.3 Instance Level Aggregation

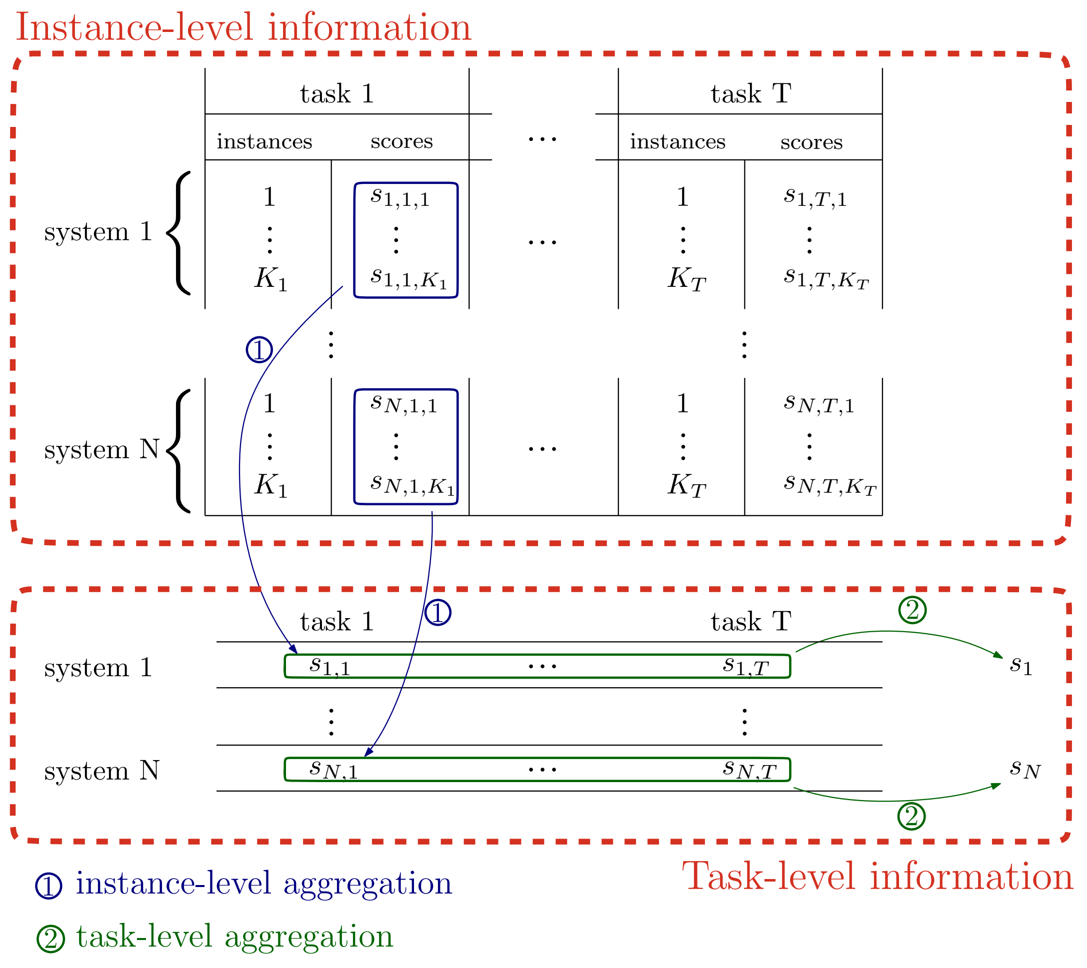
Focus on task-level ranking

Focus on task-level ranking



Initial information: $s_{n,t} \in \mathbb{R}$

Focus on task-level ranking



Initial information: $s_{n,t} \in \mathbb{R}$

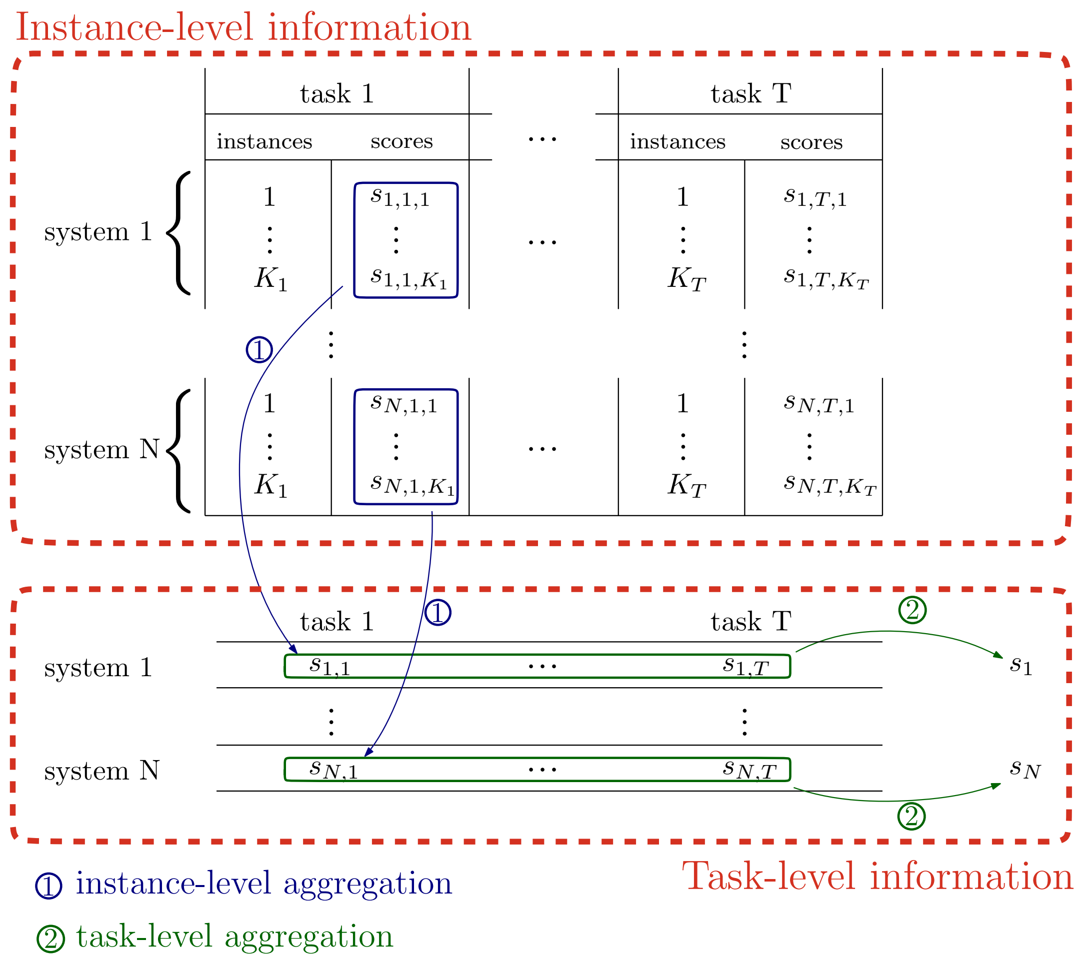
First attempt: mean-aggregation

1. **Compute aggregated scores:**

$$s_n = \sum_{t=1}^{T} s_{n,t}$$

2. **Rank systems accordingly**

Focus on task-level ranking



Initial information: $s_{n,t} \in \mathbb{R}$

First attempt: mean-aggregation

1. **Compute aggregated scores:**

$$s_n = \sum_{t=1}^{T} s_{n,t}$$

2. **Rank systems accordingly**

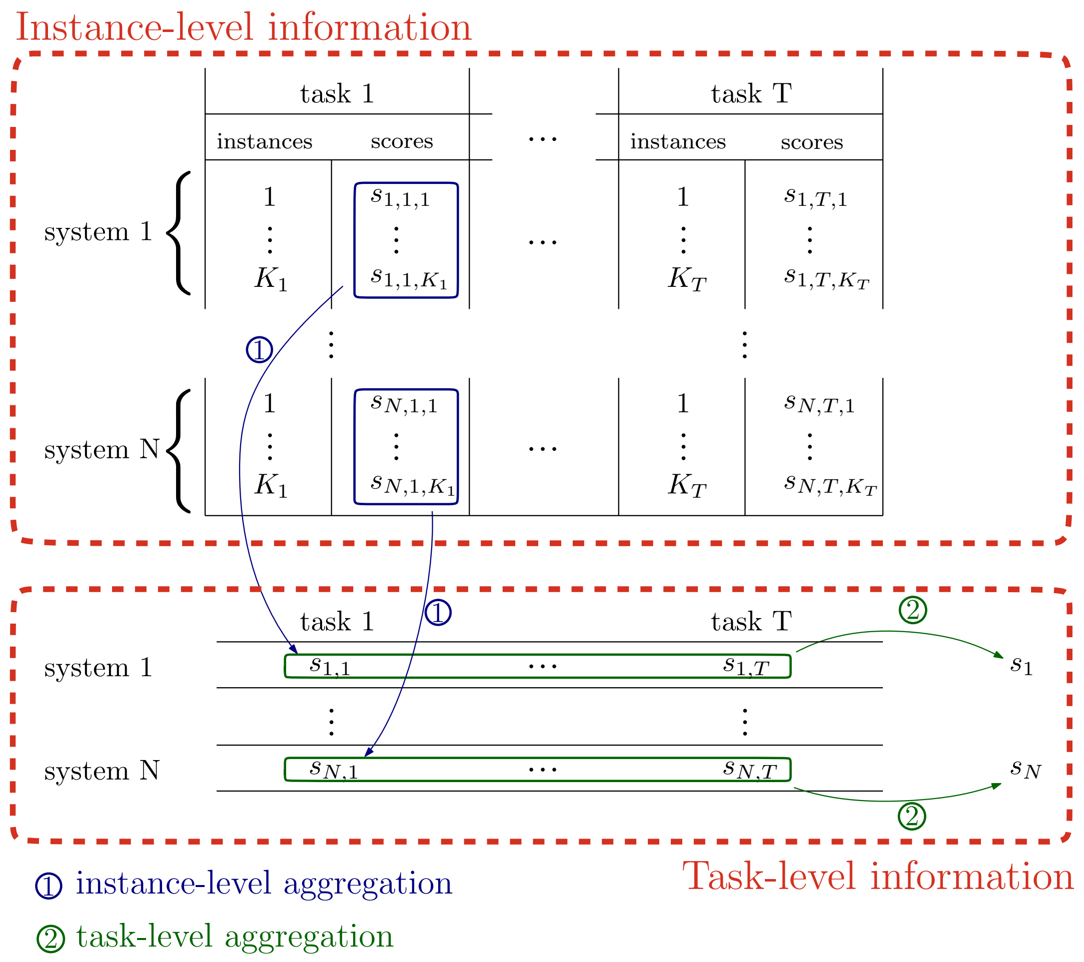
Weaknesses

1. **Scale dependent**

2. **Non-relative score**

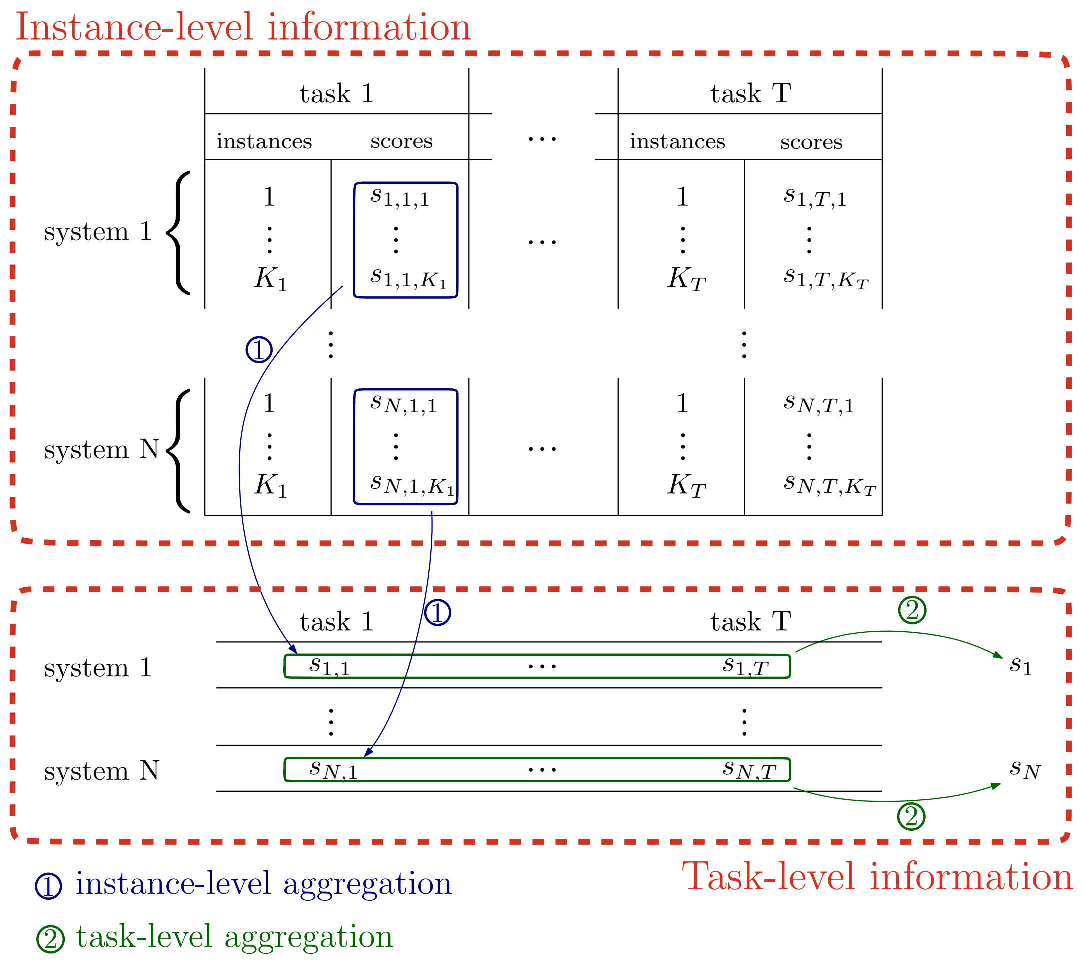
Focus on task-level ranking

Focus on task-level ranking



Initial information: $s_{n,t} \in \mathbb{R}$

Focus on task-level ranking



Initial information: $s_{n,t} \in \mathbb{R}$

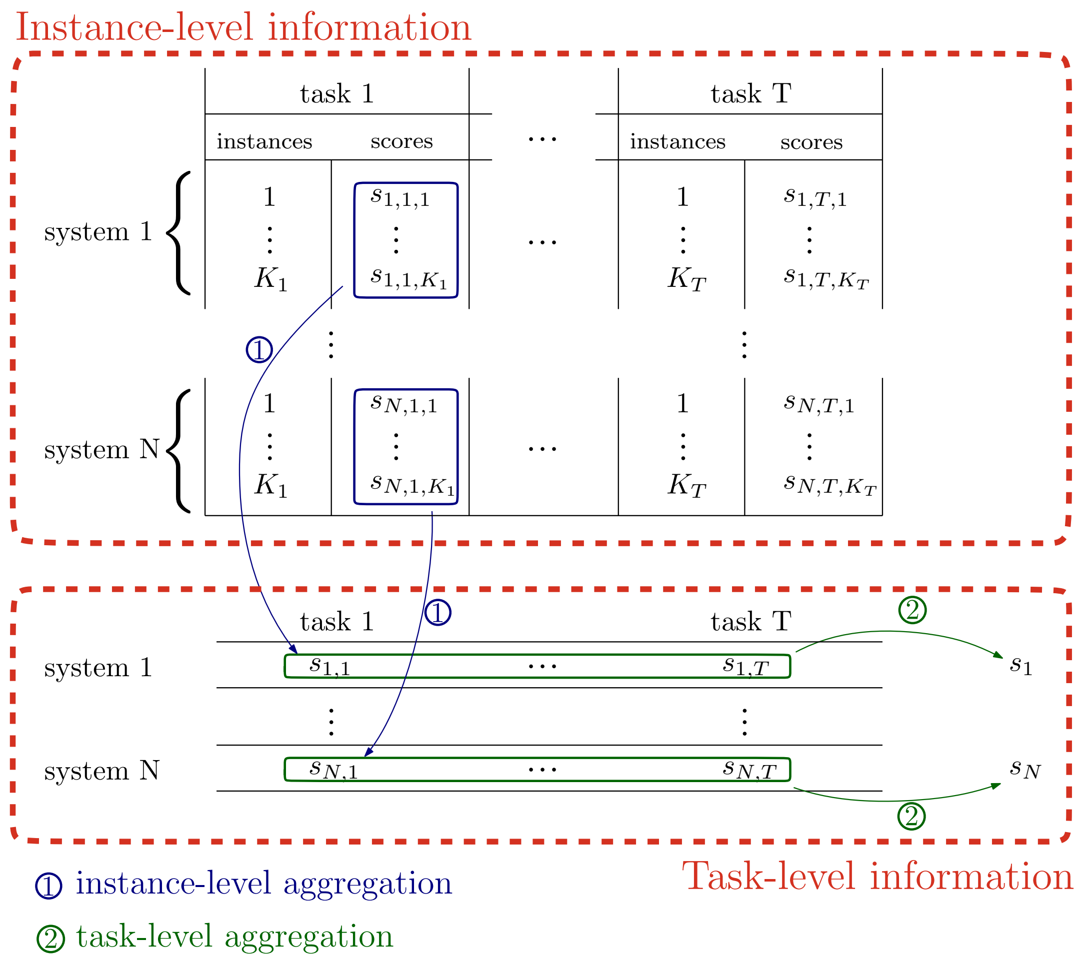
Second attempt: pairwise ranking

1. **Compute pairwise ranking:**

$$\lambda_A = \sum_{t=1}^{T} \mathbf{1}_{s_{A,t} > s_{B,t}}$$

2. **Rank A>B if and only if $\lambda_A > \lambda_B$**

Focus on task-level ranking



Initial information: $s_{n,t} \in \mathbb{R}$

Second attempt: pairwise ranking

1. **Compute pairwise ranking:**

$$\lambda_A = \sum_{t=1}^{T} \mathbf{1}_{s_{A,t} > s_{B,t}}$$

2. **Rank A>B if and only if** $\lambda_A > \lambda_B$

Weaknesses

1 Restricted to two systems

2 Can lead to paradoxes

A toy example

	task 1	task 2	task 3	task 4	task 5	task 6	sum
A	$0,3$ 3	5 3	10 1	$0,02$ 2	$1,0$ 1	$0,4$ 3	$16,72$ 13
B	$0,1$ 2	4 2	13 2	$0,01$ 1	$2,2$ 3	$0,3$ 2	$19,61$ 12
C	$0,0$ 1	3 1	15 3	$0,03$ 3	$2,0$ 2	$0,2$ 1	$20,23$ 11

mean-aggregation: $A > B > C$

pairwise ranking: $B > A, \ C > B, \ A = C$

our ranking: $C > B > A$

Our proposition: Borda's count

Our proposition: Borda's count

For every t, let σ^t be the ranking of the systems on task t :

$$\sigma^t = [\sigma_1^t, \ldots, \sigma_N^t],$$

where σ_i^t is the rank of system i.

Our proposition: Borda's count

For every t, let σ^t be the ranking of the systems on task t :

$$\sigma^t = [\sigma_1^t, \ldots, \sigma_N^t],$$

where σ_i^t is the rank of system i.

Example: $[2,1,3]$ means that system 1 is the second best, system 2 is the best and system 3 is the third best.

Our proposition: Borda's count

For every t, let σ^t be the ranking of the systems on task t :

$$\sigma^t = [\sigma_1^t, \ldots, \sigma_N^t],$$

where σ_i^t is the rank of system i.

Example: $[2,1,3]$ means that system 1 is the second best, system 2 is the best and system 3 is the third best.

1. For every system n, compute: $b_n = \sum_{t=1}^{T} \sigma_n^t$
2. Rank the systems accordingly

Why is it relevant? Elements of social choice theory

Why is it relevant? Elements of social choice theory

For every t, let σ^t be the ranking of the systems on task t :

$$\sigma^t = [\sigma_1^t, \ldots, \sigma_N^t],$$

where σ_i^t is the rank of system i.

Why is it relevant? Elements of social choice theory

For every t, let σ^t be the ranking of the systems on task t :

$$\sigma^t = [\sigma_1^t, \ldots, \sigma_N^t],$$

where σ_i^t is the rank of system i.

It is better to interpret σ^t as a permutation: $\sigma^t \in \mathfrak{S}_N$

Why is it relevant? Elements of social choice theory

For every t, let σ^t be the ranking of the systems on task t :

$$\sigma^t = [\sigma_1^t, \ldots, \sigma_N^t],$$

where σ_i^t is the rank of system i.

It is better to interpret σ^t as a permutation: $\sigma^t \in \mathfrak{S}_N$

Example: $[2,1,3]$ is the permutation that moves first object to second position, second object to first position and leaves third object in third position.

Why is it relevant? Elements of social choice theory

For every t, let σ^t be the ranking of the systems on task t :

$$\sigma^t = [\sigma_1^t, \ldots, \sigma_N^t],$$

where σ_i^t is the rank of system i.

It is better to interpret σ^t as a permutation: $\sigma^t \in \mathfrak{S}_N$

Example: $[2,1,3]$ is the permutation that moves first object to second position, second object to first position and leaves third object in third position.

$$[2,1,3] \cdot (a, b, c) = (b, a, c)$$

Why is it relevant? Elements of social choice theory

Why is it relevant? Elements of social choice theory

Each task t induces a permutation of the systems $\sigma^t \in \mathfrak{S}_N$.

Why is it relevant? Elements of social choice theory

Each task t induces a permutation of the systems $\sigma^t \in \mathfrak{S}_N$.

$\rightsquigarrow$ **New question: how to aggregate permutations?**

Why is it relevant? Elements of social choice theory

Each task t induces a permutation of the systems $\sigma^t \in \mathfrak{S}_N$.

$\rightsquigarrow$ **New question: how to aggregate permutations?**

The mean $\frac{1}{T}\sum_{t=1}^{T}\sigma^t$ makes no sense!

Why is it relevant? Elements of social choice theory

Each task t induces a permutation of the systems $\sigma^t \in \mathfrak{S}_N$.

$\rightsquigarrow$ **New question: how to aggregate permutations?**

The mean $\frac{1}{T}\sum_{t=1}^{T} \sigma^t$ makes no sense!

Solution: define a distance d on the permutation group, and find a permutation σ^* that minimizes the sum of distances:

$$\sigma^* \in \operatorname*{argmin}_{\sigma \in \mathfrak{S}_N} \sum_{t=1}^{T} d(\sigma, \sigma^t)$$

Why is it relevant? Elements of social choice theory

Why is it relevant? Elements of social choice theory

⇝ **How to aggregate permutations?**

$$\sigma^* \in \operatorname*{argmin}_{\sigma \in \mathfrak{S}_N} \sum_{t=1}^{T} d(\sigma, \sigma^t)$$

Why is it relevant? Elements of social choice theory

⇝ **How to aggregate permutations?**

$$\sigma^* \in \operatorname{argmin}_{\sigma \in \mathfrak{S}_N} \sum_{t=1}^{T} d(\sigma, \sigma^t)$$

When d is the Kendall distance that counts the number of inversions, σ^* is called a Kemeny consensus.

Why is it relevant? Elements of social choice theory

⇝ How to aggregate permutations?

$$\sigma^* \in \underset{\sigma \in \mathfrak{S}_N}{\operatorname{argmin}} \sum_{t=1}^{T} d(\sigma, \sigma^t)$$

When d is the Kendall distance that counts the number of inversions, σ^* is called a Kemeny consensus.

It is the only aggregation of permutations procedures that satisfies three natural axioms.

- Neutrality
- Consistency
- Condorcet Criterion

Why is it relevant? Elements of social choice theory

⇝ How to aggregate permutations?

$$\sigma^* \in \underset{\sigma \in \mathfrak{S}_N}{\operatorname{argmin}} \sum_{t=1}^{T} d(\sigma, \sigma^t)$$

When d is the Kendall distance that counts the number of inversions, σ^* is called a Kemeny consensus.

It is the only aggregation of permutations procedures that satisfies three natural axioms.

- Neutrality
- Consistency
- Condorcet Criterion

BUT: NP-Hard problem!

Why is it relevant? Elements of social choice theory

⇝ **How to aggregate permutations?**

$$\sigma^* \in \underset{\sigma \in \mathfrak{S}_N}{\operatorname{argmin}} \sum_{t=1}^{T} d(\sigma, \sigma^t)$$

When d is the Kendall distance that counts the number of inversions, σ^* is called a Kemeny consensus.

It is the only aggregation of permutations procedures that satisfies three natural axioms.

- Neutrality
- Consistency
- Condorcet Criterion

BUT: NP-Hard problem!

Relaxation of the problem: Borda count!

+ 2-approximation
+ Small complexity
+ Simple interpretation

Numerical Results

Numerical Results

Ranking Analysis

GLUE			XTREM		
σ^*	Team	σ^{mean}	σ^*	Team	σ^{mean}
0 (1430)	Ms Alex	0 (88.6)	0 (55)	ULR	0 (83.2)
1 (1405)	ERNIE	1 (88.0)	1 (50)	CoFe	1 (82.6)
2 (1397)	DEBERTA	2 (87.9)	2 (44)	InfoLXL	3 (80.6)
3 (1391)	AliceMind	3 (87.8)	3 (42)	VECO	4 (80.3)
4 (1375)	PING-AH	5 (87.6)	4 (35)	Unicoder	5 (79.4)
5 (1362)	HFL	4 (87.7)	5 (34)	PolyGlot	2 (80.6)
6 (1361)	T5	6 (87.5)	6 (31)	ULR-v2	6 (79.4)
7 (1358)	DIRL	10 (86.7)	7 (29)	HiCTL	8 (79.1)
8 (1331)	Zihan	7 (87.6)	8 (29)	Ernie	7 (79.1)
9 (1316)	ELECTRA	11 (86.7)	9 (21)	Anony	10 (78.3)

Numerical Results

Ranking Analysis

GLUE			XTREM		
σ^*	Team	σ^{mean}	σ^*	Team	σ^{mean}
0 (1430)	Ms Alex	0 (88.6)	0 (55)	ULR	0 (83.2)
1 (1405)	ERNIE	1 (88.0)	1 (50)	CoFe	1 (82.6)
2 (1397)	DEBERTA	2 (87.9)	2 (44)	InfoLXL	3 (80.6)
3 (1391)	AliceMind	3 (87.8)	3 (42)	VECO	4 (80.3)
4 (1375)	PING-AH	5 (87.6)	4 (35)	Unicoder	5 (79.4)
5 (1362)	HFL	4 (87.7)	5 (34)	PolyGlot	2 (80.6)
6 (1361)	T5	6 (87.5)	6 (31)	ULR-v2	6 (79.4)
7 (1358)	DIRL	10 (86.7)	7 (29)	HiCTL	8 (79.1)
8 (1331)	Zihan	7 (87.6)	8 (29)	Ernie	7 (79.1)
9 (1316)	ELECTRA	11 (86.7)	9 (21)	Anony	10 (78.3)

Numerical Results

Ranking Analysis

GLUE			XTREM		
σ^*	Team	σ^{mean}	σ^*	Team	σ^{mean}
0 (1430)	Ms Alex	0 (88.6)	0 (55)	ULR	0 (83.2)
1 (1405)	ERNIE	1 (88.0)	1 (50)	CoFe	1 (82.6)
2 (1397)	DEBERTA	2 (87.9)	2 (44)	InfoLXL	3 (80.6)
3 (1391)	AliceMind	3 (87.8)	3 (42)	VECO	4 (80.3)
4 (1375)	PING-AH	5 (87.6)	4 (35)	Unicoder	5 (79.4)
5 (1362)	HFL	4 (87.7)	5 (34)	PolyGlot	2 (80.6)
6 (1361)	T5	6 (87.5)	6 (31)	ULR-v2	6 (79.4)
7 (1358)	DIRL	10 (86.7)	7 (29)	HiCTL	8 (79.1)
8 (1331)	Zihan	7 (87.6)	8 (29)	Ernie	7 (79.1)
9 (1316)	ELECTRA	11 (86.7)	9 (21)	Anony	10 (78.3)

Numerical Results

Ranking Analysis

GLUE			XTREM		
σ^*	Team	σ^{mean}	σ^*	Team	σ^{mean}
0 (1430)	Ms Alex	0 (88.6)	0 (55)	ULR	0 (83.2)
1 (1405)	ERNIE	1 (88.0)	1 (50)	CoFe	1 (82.6)
2 (1397)	DEBERTA	2 (87.9)	2 (44)	InfoLXL	3 (80.6)
3 (1391)	AliceMind	3 (87.8)	3 (42)	VECO	4 (80.3)
4 (1375)	PING-AH	5 (87.6)	4 (35)	Unicoder	5 (79.4)
5 (1362)	HFL	4 (87.7)	5 (34)	PolyGlot	2 (80.6)
6 (1361)	T5	6 (87.5)	6 (31)	ULR-v2	6 (79.4)
7 (1358)	DIRL	10 (86.7)	7 (29)	HiCTL	8 (79.1)
8 (1331)	Zihan	7 (87.6)	8 (29)	Ernie	7 (79.1)
9 (1316)	ELECTRA	11 (86.7)	9 (21)	Anony	10 (78.3)

Aggregation procedure matters a lot!

Numerical Results

Ranking Analysis

GLUE			XTREM		
σ^*	Team	σ^{mean}	σ^*	Team	σ^{mean}
0 (1430)	Ms Alex	0 (88.6)	0 (55)	ULR	0 (83.2)
1 (1405)	ERNIE	1 (88.0)	1 (50)	CoFe	1 (82.6)
2 (1397)	DEBERTA	2 (87.9)	2 (44)	InfoLXL	3 (80.6)
3 (1391)	AliceMind	3 (87.8)	3 (42)	VECO	4 (80.3)
4 (1375)	PING-AH	5 (87.6)	4 (35)	Unicoder	5 (79.4)
5 (1362)	HFL	4 (87.7)	5 (34)	PolyGlot	2 (80.6)
6 (1361)	T5	6 (87.5)	6 (31)	ULR-v2	6 (79.4)
7 (1358)	DIRL	10 (86.7)	7 (29)	HiCTL	8 (79.1)
8 (1331)	Zihan	7 (87.6)	8 (29)	Ernie	7 (79.1)
9 (1316)	ELECTRA	11 (86.7)	9 (21)	Anony	10 (78.3)

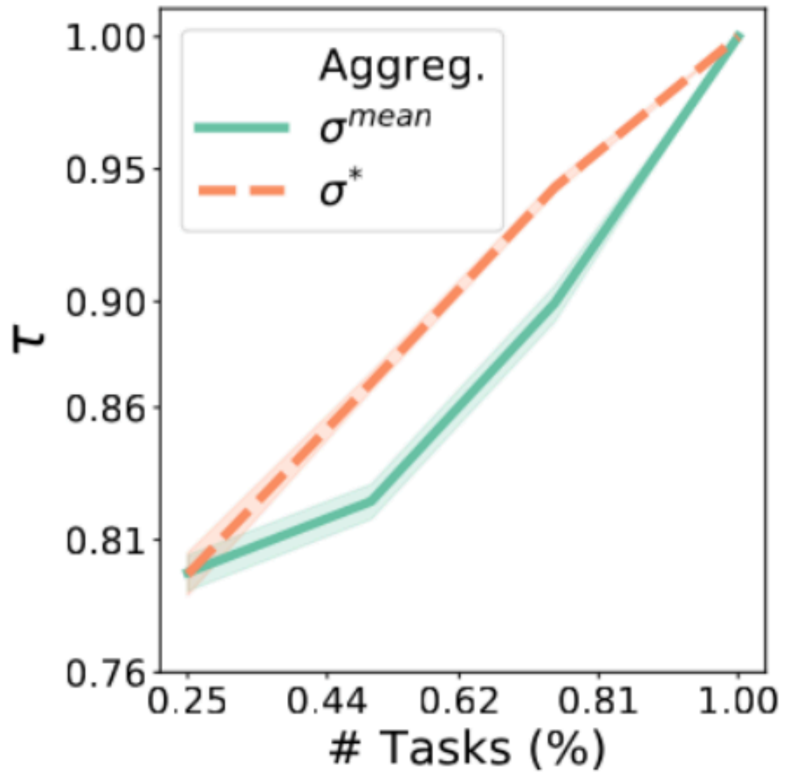
Aggregation procedure matters a lot!

Robustness Analysis

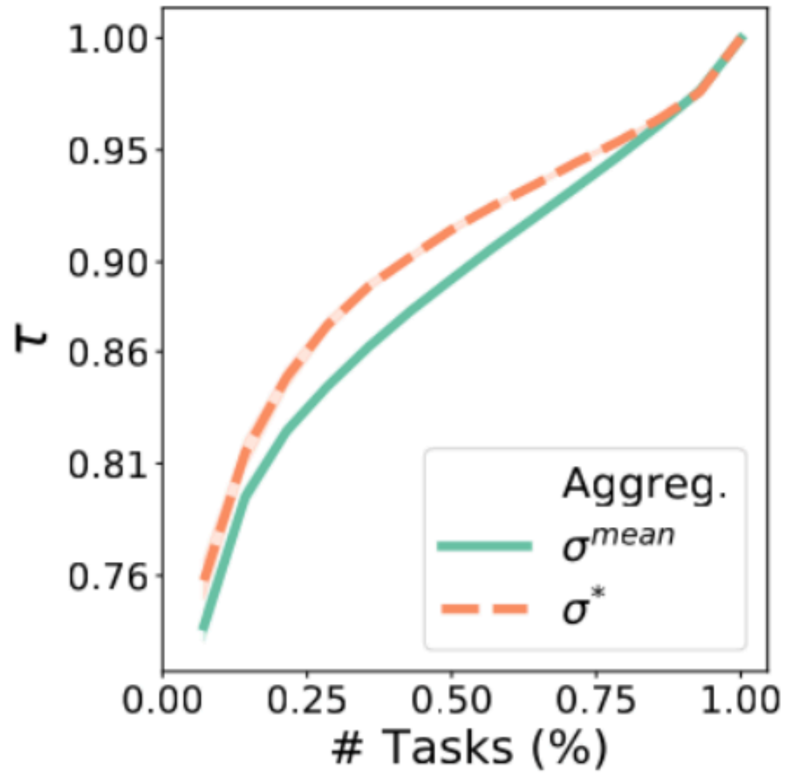
Setting:

For an increasing % of task, compute the Kendall's tau correlation coefficient between the obtained ranking and the one obtained with all the tasks

Higher is better



(f) EXTREM



(a) GLUE

Numerical Results

Ranking Analysis

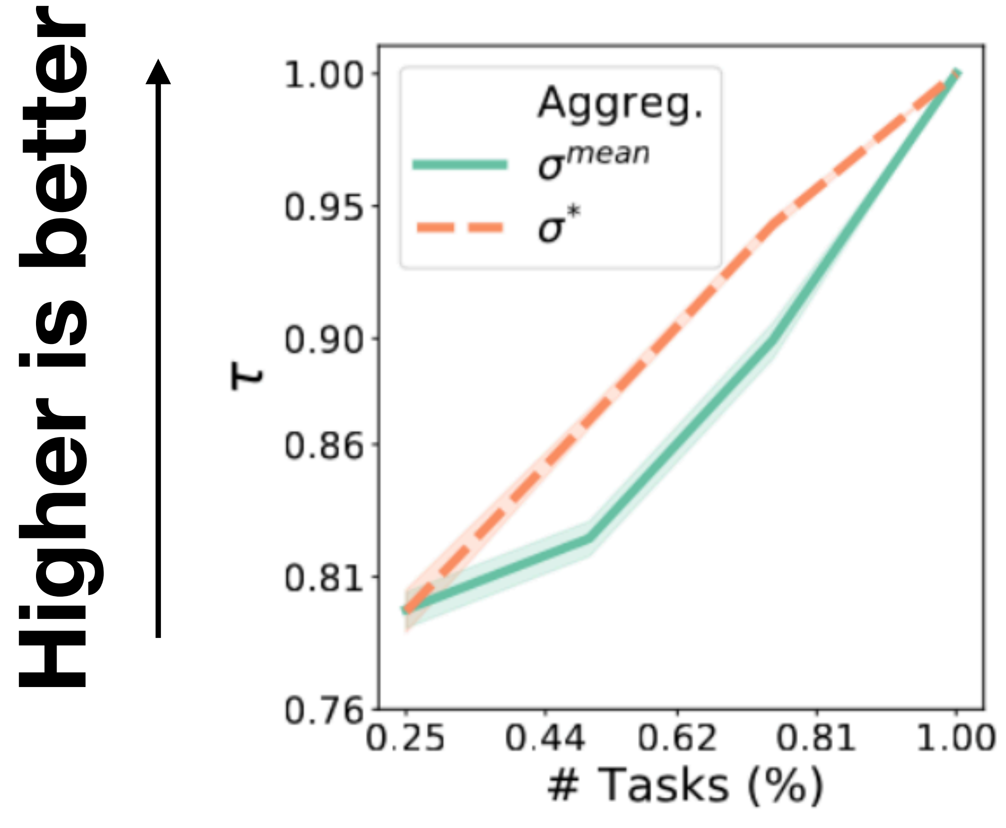
GLUE			XTREM		
σ^*	Team	σ^{mean}	σ^*	Team	σ^{mean}
0 (1430)	Ms Alex	0 (88.6)	0 (55)	ULR	0 (83.2)
1 (1405)	ERNIE	1 (88.0)	1 (50)	CoFe	1 (82.6)
2 (1397)	DEBERTA	2 (87.9)	2 (44)	InfoLXL	3 (80.6)
3 (1391)	AliceMind	3 (87.8)	3 (42)	VECO	4 (80.3)
4 (1375)	PING-AH	5 (87.6)	4 (35)	Unicoder	5 (79.4)
5 (1362)	HFL	4 (87.7)	5 (34)	PolyGlot	2 (80.6)
6 (1361)	T5	6 (87.5)	6 (31)	ULR-v2	6 (79.4)
7 (1358)	DIRL	10 (86.7)	7 (29)	HiCTL	8 (79.1)
8 (1331)	Zihan	7 (87.6)	8 (29)	Ernie	7 (79.1)
9 (1316)	ELECTRA	11 (86.7)	9 (21)	Anony	10 (78.3)

Aggregation procedure matters a lot!

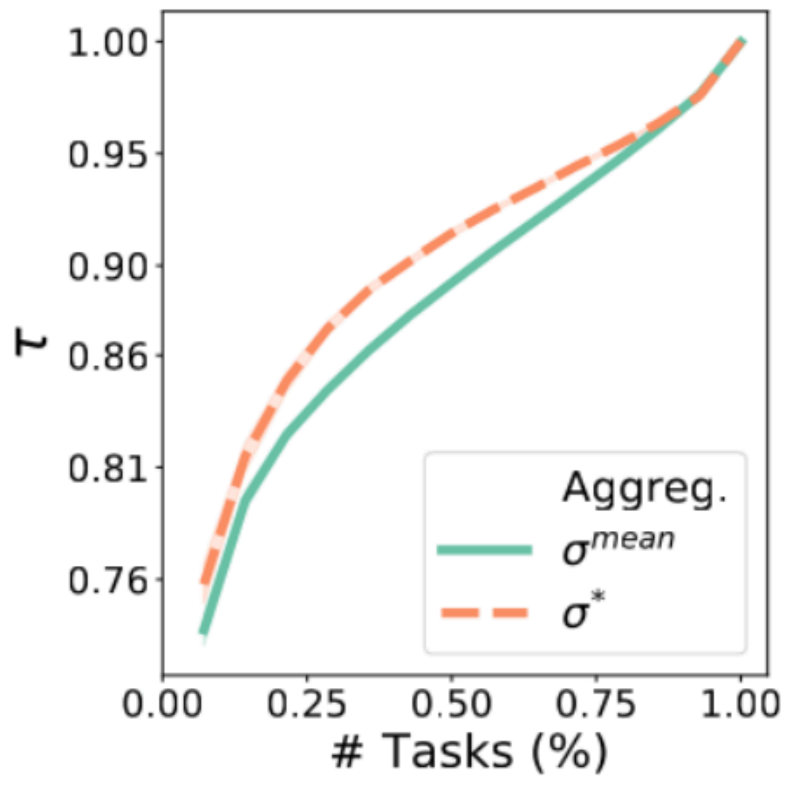
Robustness Analysis

Setting:

For an increasing % of task, compute the Kendall's tau correlation coefficient between the obtained ranking and the one obtained with all the tasks



(f) EXTREM



(a) GLUE

Relying on Borda count is more reliable

2. How to aggregate several metrics?

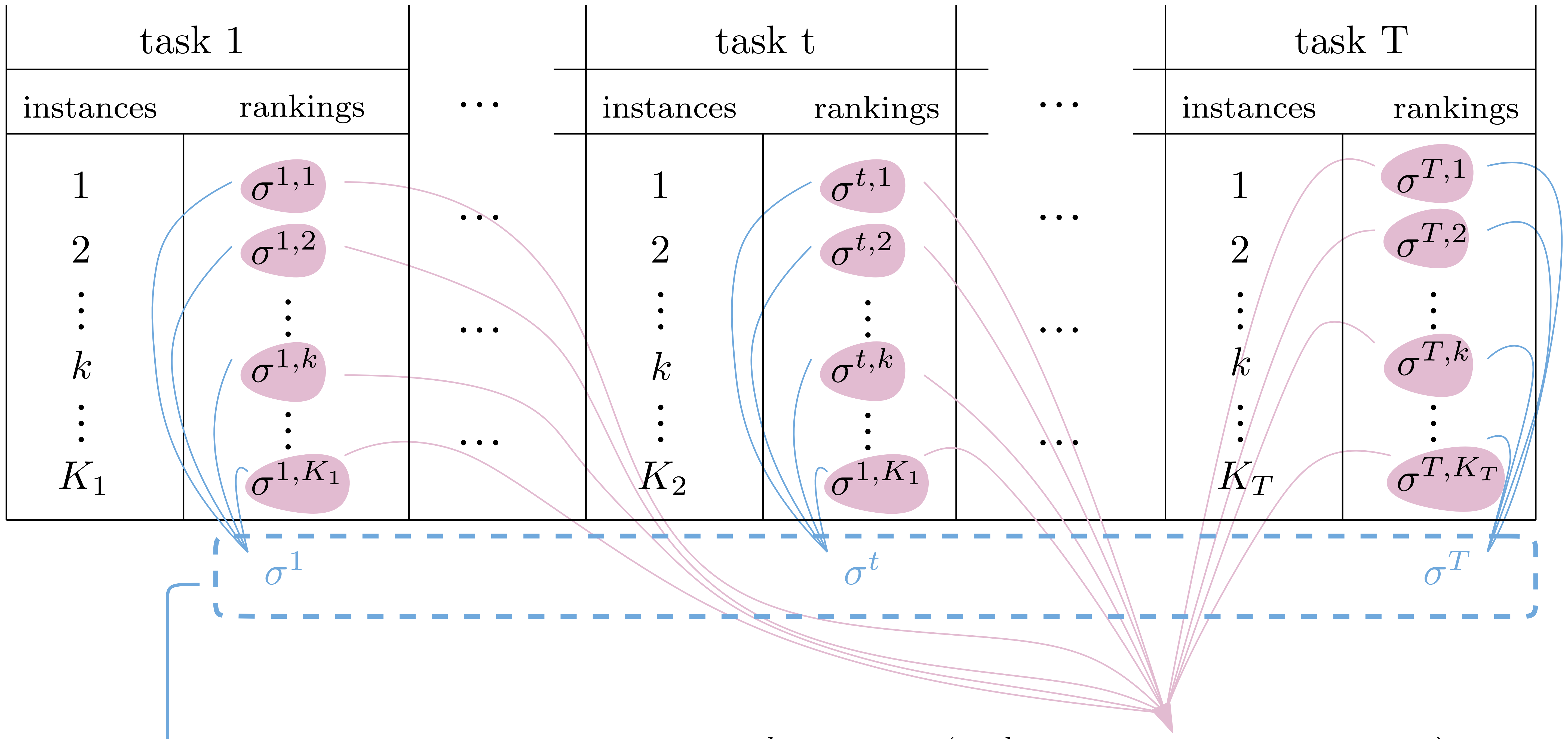
1.1 Framework

1.2 Task Level Aggregation

1.3 Instance Level Aggregation

What about instance-level aggregation?

What about instance-level aggregation?

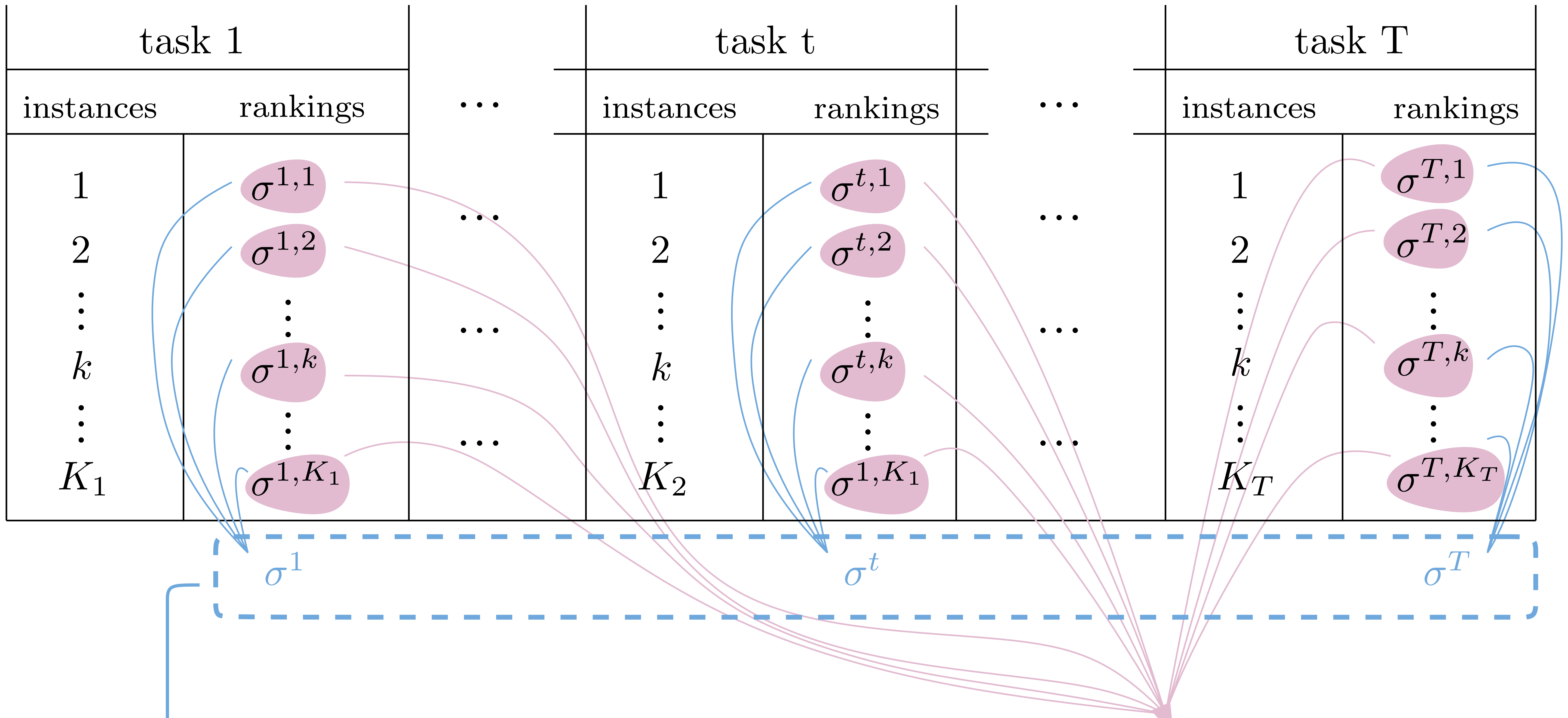


One level aggregation: $\sigma^l = \mathrm{Borda}\big(\sigma^{t,k},\ 1 \leq t \leq T,\ 1 \leq k \leq K_t\big)$

Two level aggregation: $\forall 1 \leq t \leq T, \quad \sigma^t = \mathrm{Borda}\big(\sigma^{t,k},\ 1 \leq k \leq K_t\big)$

$\sigma^{2l} = \mathrm{Borda}\big(\sigma^t,\ 1 \leq t \leq T\big)$

What about instance-level aggregation?



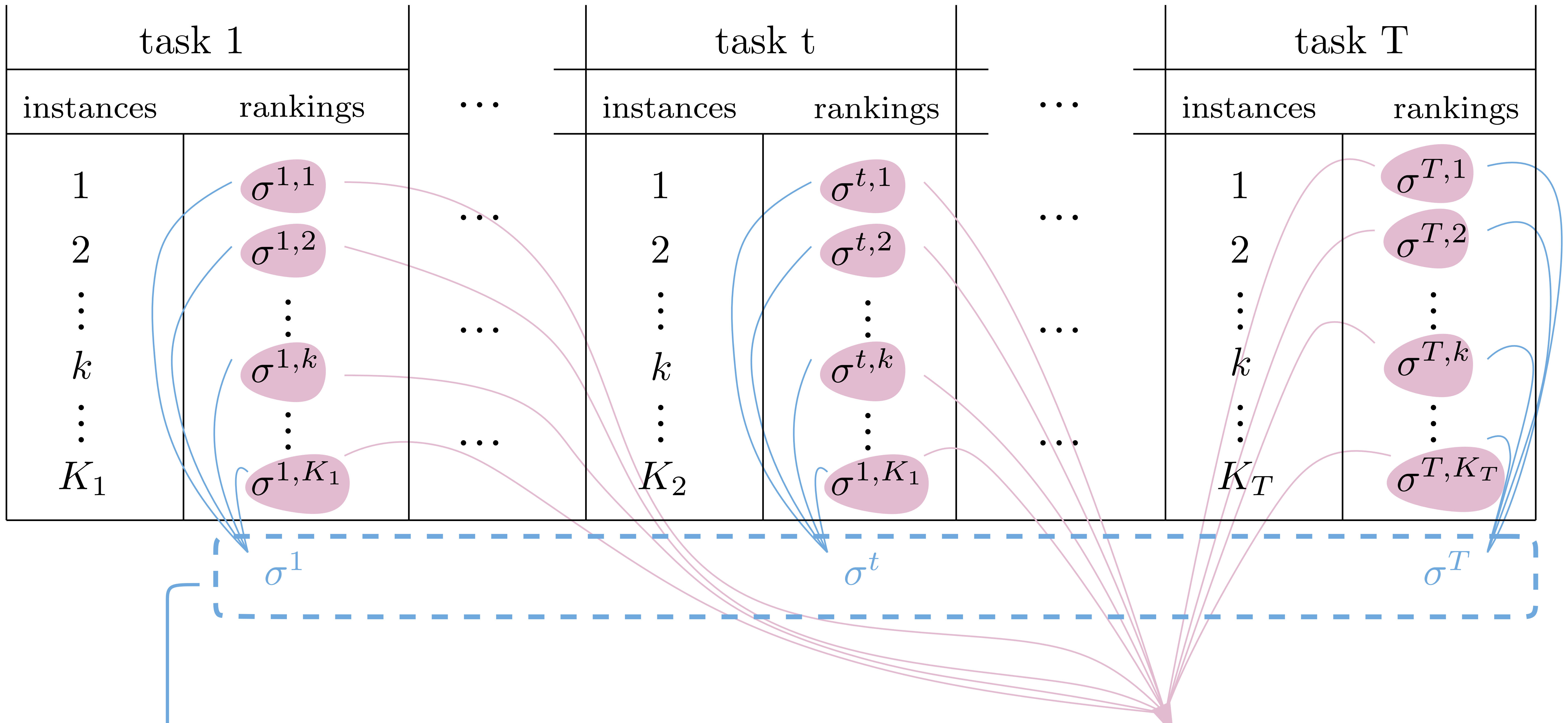
For every n, every t and every k, access to the aggregated performance of system n on instance k of task t

One level aggregation: $\sigma^l = \mathrm{Borda}\big(\sigma^{t,k},\ 1 \leq t \leq T,\ 1 \leq k \leq K_t\big)$

Two level aggregation: $\forall 1 \leq t \leq T, \quad \sigma^t = \mathrm{Borda}\big(\sigma^{t,k},\ 1 \leq k \leq K_t\big)$

$\sigma^{2l} = \mathrm{Borda}\big(\sigma^t,\ 1 \leq t \leq T\big)$

What about instance-level aggregation?



One level aggregation: $\sigma^l = \text{Borda}\big(\sigma^{t,k},\ 1 \le t \le T,\ 1 \le k \le K_t\big)$

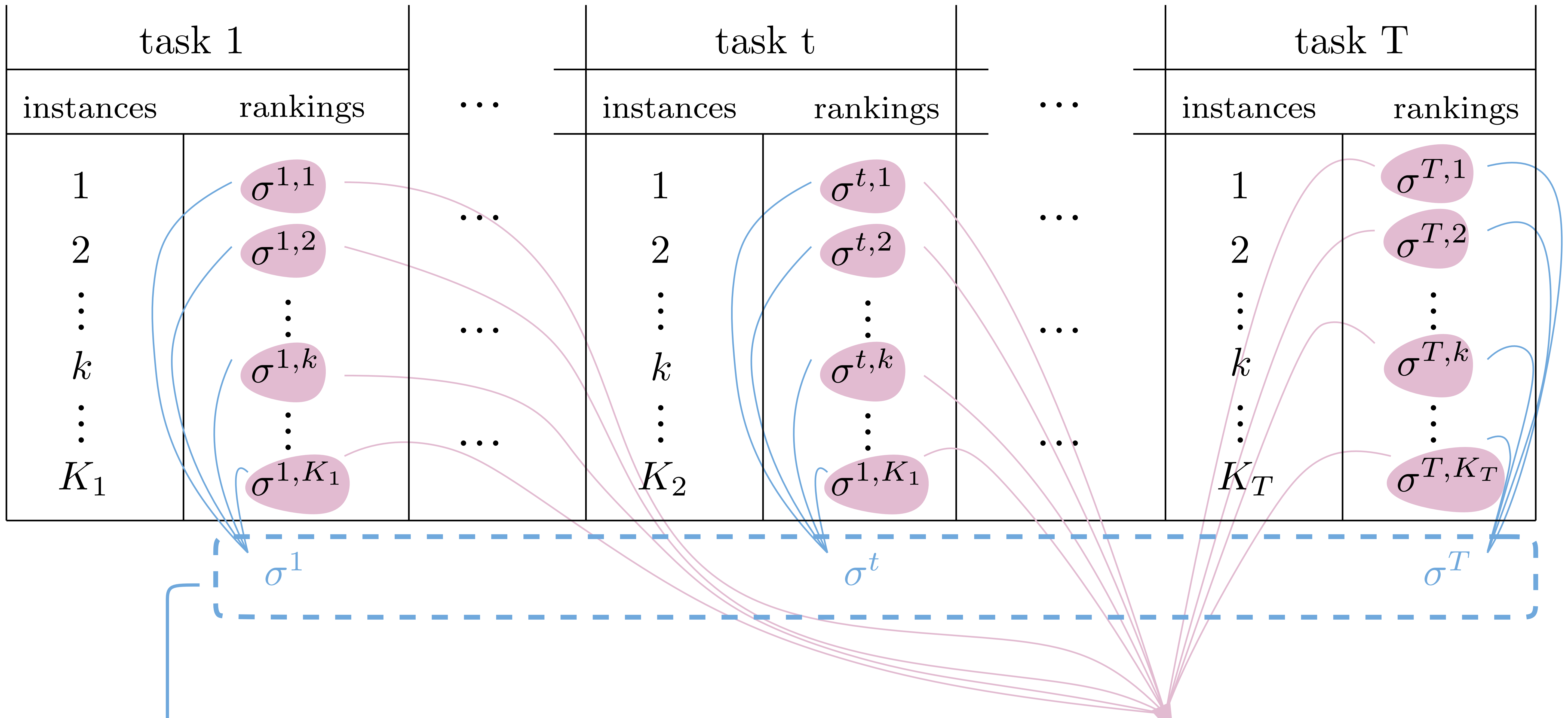
Two level aggregation: $\forall 1 \le t \le T, \quad \sigma^t = \text{Borda}\big(\sigma^{t,k},\ 1 \le k \le K_t\big)$

$\sigma^{2l} = \text{Borda}\big(\sigma^t,\ 1 \le t \le T\big)$

For every n, every t and every k, access to the aggregated performance of system n on instance k of task t

$$s_{n,t,k} \in \mathbb{R}$$

What about instance-level aggregation?



One level aggregation: $\sigma^l = \text{Borda}\big(\sigma^{t,k},\ 1 \le t \le T,\ 1 \le k \le K_t\big)$

Two level aggregation: $\forall 1 \le t \le T, \quad \sigma^t = \text{Borda}\big(\sigma^{t,k},\ 1 \le k \le K_t\big)$

$\sigma^{2l} = \text{Borda}\big(\sigma^t,\ 1 \le t \le T\big)$

For every n, every t and every k, access to the aggregated performance of system n on instance k of task t

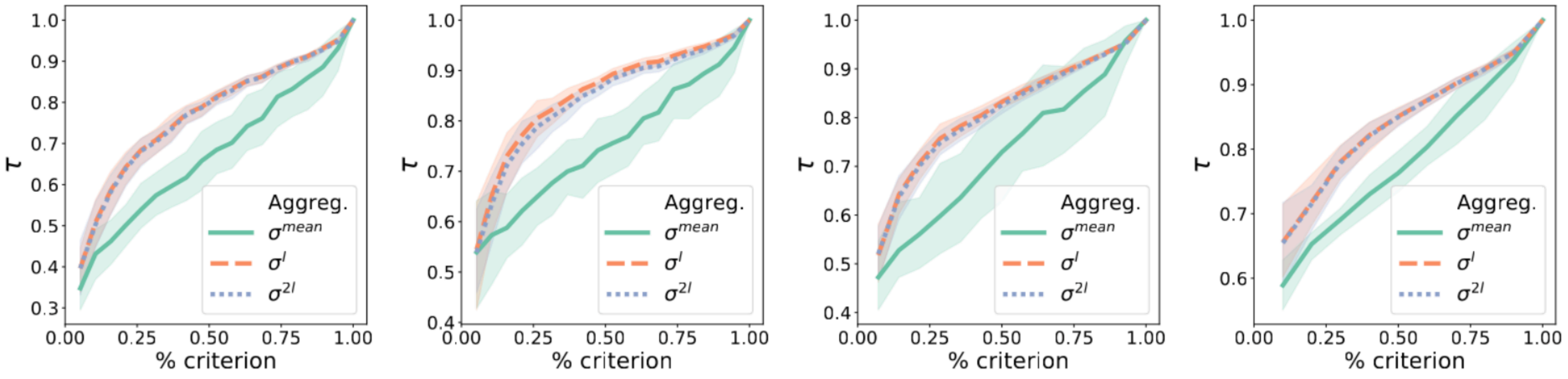
$$s_{n,t,k} \in \mathbb{R}$$

Goal: find an aggregation procedure that orders the systems.

Numerical Results

Numerical Results

Robustness Analysis

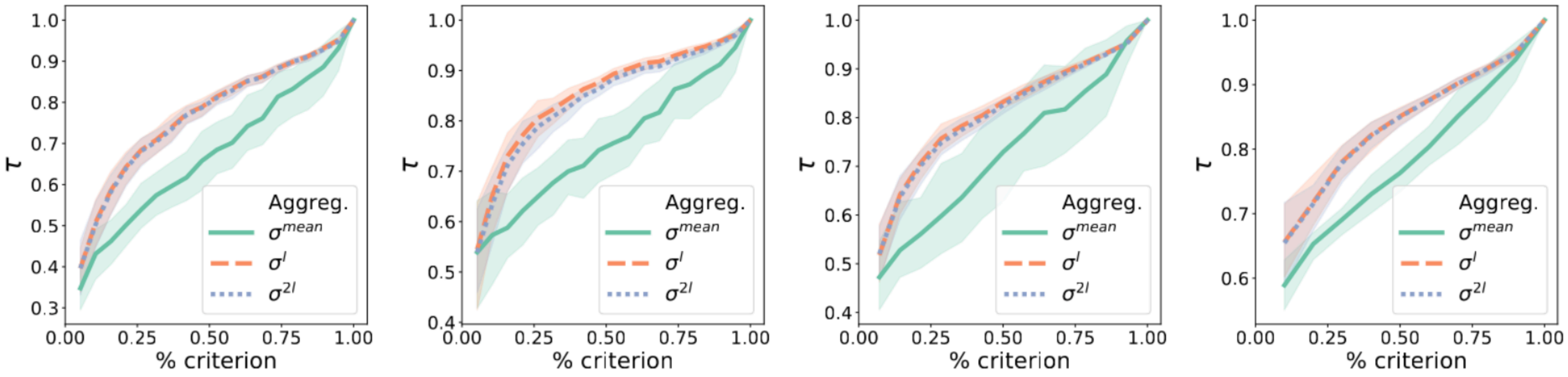


(b) Persona Chat (c) Topic Chat (d) FLICKR (e) MLQE

Numerical Results

Robustness Analysis

Relying on Borda count is more reliable. An 1 or 2 level are equivalents.

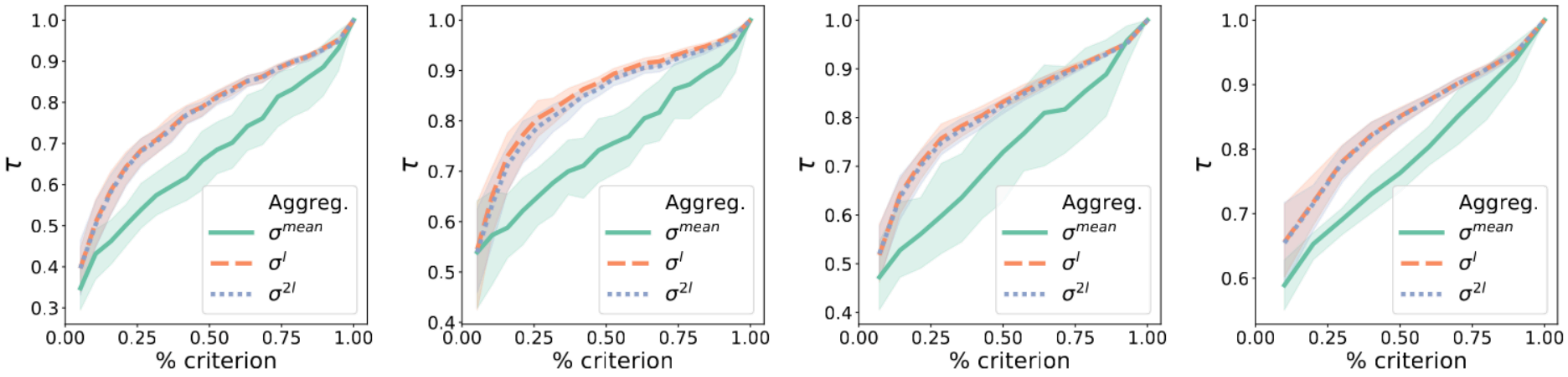


(b) Persona Chat (c) Topic Chat (d) FLICKR (e) MLQE

Numerical Results

Robustness Analysis

Relying on Borda count is more reliable. An 1 or 2 level are equivalents.



(b) Persona Chat (c) Topic Chat (d) FLICKR (e) MLQE

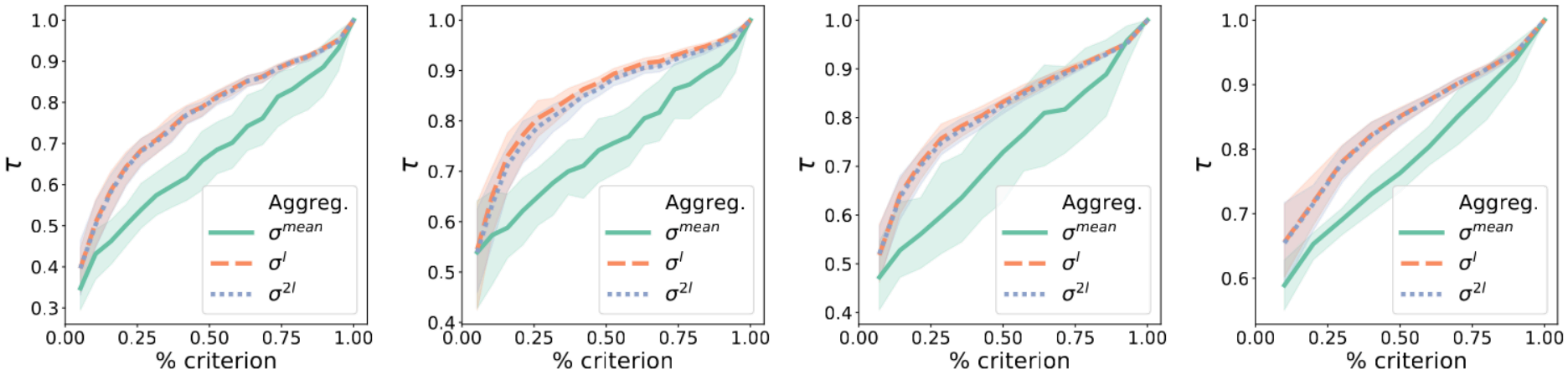
Ranking Correlation

	PC	TC	FLI.	MLQE
$\tau(\sigma^{l}, \sigma^{2l})$	-0.08	-0.01	0	-0.03
$\tau(\sigma^{mean}, \sigma^{2l})$	0.32	0.27	0.29	0.01
$\tau(\sigma^{mean}, \sigma^{l})$	-0.10	-0.15	-0.04	0.00
RSUM	SEVAL	TAC08	TAC09	TAC11
0.04	0.14	0.28	0.06	-0.06
0.07	0.52	0.32	0.37	0.37
0	0.10	0.23	0.19	0.07

Numerical Results

Robustness Analysis

Relying on Borda count is more reliable. An 1 or 2 level are equivalents.



(b) Persona Chat (c) Topic Chat (d) FLICKR (e) MLQE

Ranking Correlation

Aggregation procedure matters a lot!

σ^l disagrees from σ^{2l} and σ^{mean} both on top systems and on their orders.

σ^{2l} and σ^{mean} select similar systems but rank them differently.

	PC	TC	FLI.	MLQE
$\tau(\sigma^l, \sigma^{2l})$	-0.08	-0.01	0	-0.03
$\tau(\sigma^{mean}, \sigma^{2l})$	0.32	0.27	0.29	0.01
$\tau(\sigma^{mean}, \sigma^{l})$	-0.10	-0.15	-0.04	0.00
RSUM	SEVAL	TAC08	TAC09	TAC11
0.04	0.14	0.28	0.06	-0.06
0.07	0.52	0.32	0.37	0.37
0	0.10	0.23	0.19	0.07

3. Conclusions