Recent Evaluation Metrics for Text Generation

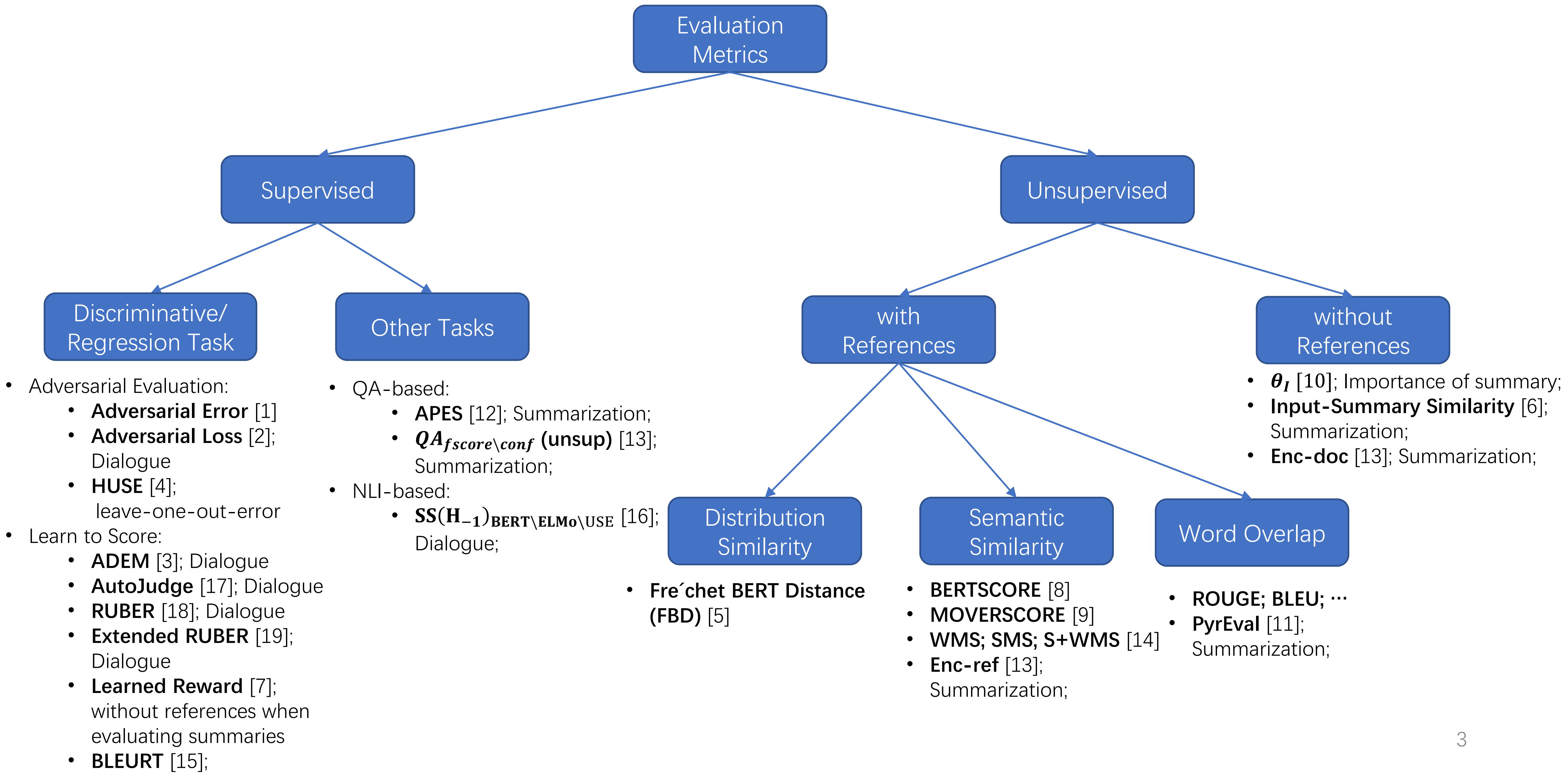
Presenter: Wang Chen

Mentor: Piji Li

Outline

- Brief Taxonomy
- Papers to Read:
  - MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance; *EMNLP-2019*
  - A Simple Theoretical Model of Importance for Summarization; *ACL-2019*
  - RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems; *AAAI-2018*
  - Better Automatic Evaluation of Open-Domain Dialogue Systems with Contextualized Embeddings; *NAACL-WS-2019*
- Key Ideas of Other Metrics
- Conclusions

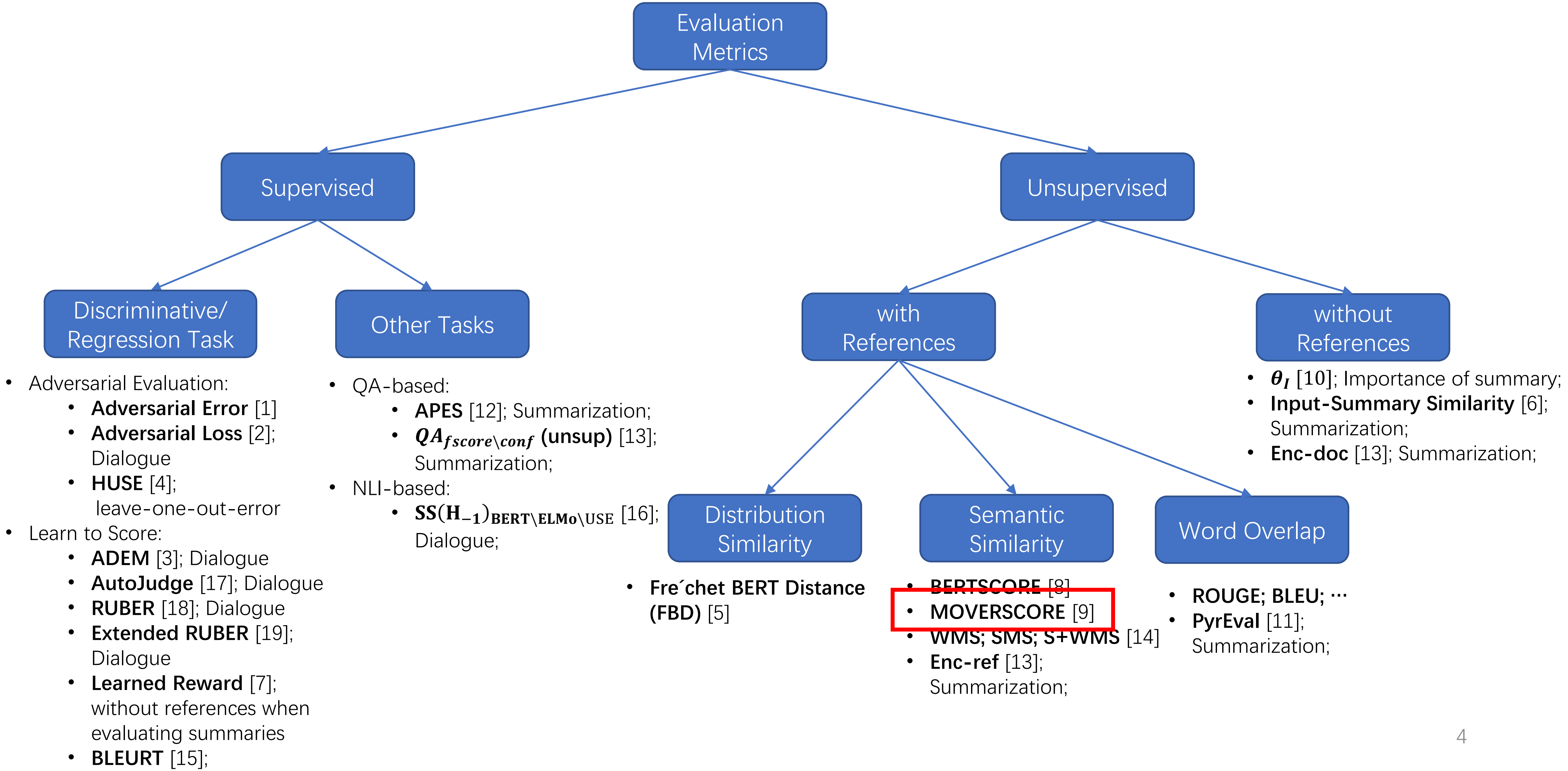
Brief Taxonomy



Evaluation Metrics

- Supervised
  - Discriminative/ Regression Task
    - Adversarial Evaluation:
      - **Adversarial Error** [1]
      - **Adversarial Loss** [2]; Dialogue
      - **HUSE** [4]; leave-one-out-error
    - Learn to Score:
      - **ADEM** [3]; Dialogue
      - **AutoJudge** [17]; Dialogue
      - **RUBER** [18]; Dialogue
      - **Extended RUBER** [19]; Dialogue
      - **Learned Reward** [7]; without references when evaluating summaries
      - **BLEURT** [15];
  - Other Tasks
    - QA-based:
      - **APES** [12]; Summarization;
      - $\boldsymbol{QA_{fscore\backslash conf}}$ **(unsup)** [13]; Summarization;
    - NLI-based:
      - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;
- Unsupervised
  - with References
    - Distribution Similarity
      - **Fre´chet BERT Distance (FBD)** [5]
    - Semantic Similarity
      - **BERTSCORE** [8]
      - **MOVERSCORE** [9]
      - **WMS; SMS; S+WMS** [14]
      - **Enc-ref** [13]; Summarization;
    - Word Overlap
      - **ROUGE; BLEU; …**
      - **PyrEval** [11]; Summarization;
  - without References
    - $\boldsymbol{\theta_I}$ [10]; Importance of summary;
    - **Input-Summary Similarity** [6]; Summarization;
    - **Enc-doc** [13]; Summarization;

Brief Taxonomy



Evaluation Metrics

Supervised

Unsupervised

Discriminative/ Regression Task

- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

Other Tasks

- QA-based:
  - **APES** [12]; Summarization;
  - $\boldsymbol{QA_{fscore\backslash conf}}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

with References

Distribution Similarity

- **Fre´chet BERT Distance (FBD)** [5]

Semantic Similarity

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

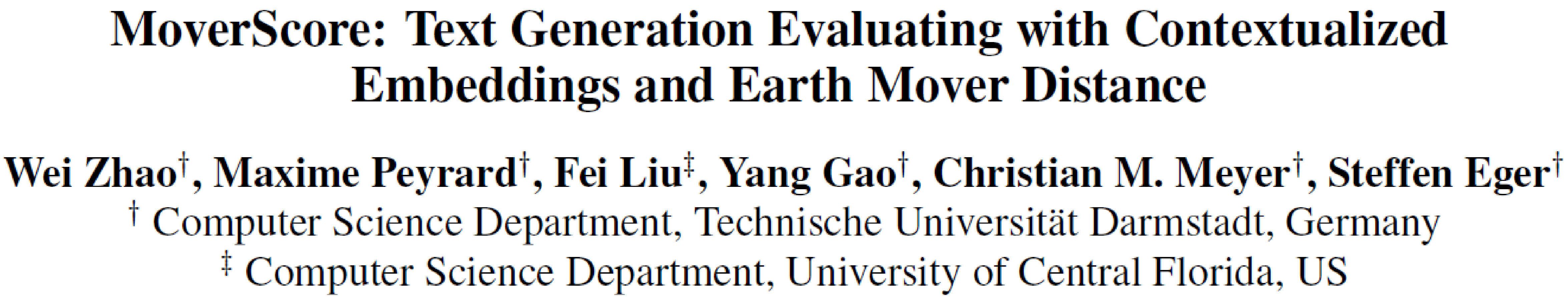
Word Overlap

- **ROUGE; BLEU; ···**
- **PyrEval** [11]; Summarization;

without References

- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

MoverScore-Title & Authors



MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance

Wei Zhao[†], Maxime Peyrard[†], Fei Liu[‡], Yang Gao[†], Christian M. Meyer[†], Steffen Eger[†]
[†] Computer Science Department, Technische Universität Darmstadt, Germany
[‡] Computer Science Department, University of Central Florida, US

MoverScore-Introduction

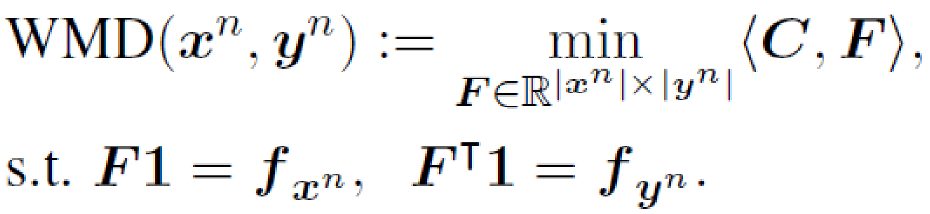
- **Motivation:** A desirable metric compares system output against references based on their semantics rather than surface forms. Distinct surface forms may convey the same meaning.

- **Method:** They investigate the effectiveness of a spectrum of distributional semantic representations to encode system and reference texts, allowing them to be compared for semantic similarity by quantifying the semantic distance.
  - BERT + Word/Sent Mover's Distance

- **Contributions:**
  1. formulate the problem of evaluating generation systems as measuring the semantic distance
  2. investigate the effectiveness of existing contextualized representations and Earth Mover's Distance
  3. outperforms or performs comparably to strong baselines on four text generation tasks including summarization, machine translation, image captioning, and data-to-text generation

MoverScore-Main Idea

- The semantic distance is computed based on the Word Mover’s Distance (WMD).



$$\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \langle \boldsymbol{C}, \boldsymbol{F} \rangle,$$

$$\text{s.t. } \boldsymbol{F}\boldsymbol{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^\intercal \boldsymbol{1} = \boldsymbol{f}_{\boldsymbol{y}^n}.$$

- System prediction $\boldsymbol{x} = (x_1, \dots, x_m)$ is a sentence viewed as a sequence of words. Reference $\boldsymbol{y}$ is also a word sequence.

MoverScore-Main Idea

- The semantic distance is computed based on the Word Mover’s Distance (WMD).

$$\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \langle \boldsymbol{C}, \boldsymbol{F} \rangle,$$

$$\text{s.t. } \boldsymbol{F}\boldsymbol{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^{\intercal}\boldsymbol{1} = \boldsymbol{f}_{\boldsymbol{y}^n}.$$

- System prediction $\boldsymbol{x} = (x_1, \dots, x_m)$ is a sentence viewed as a sequence of words. Reference $\boldsymbol{y}$ is also a word sequence.

- $\boldsymbol{x}^n$ ($\boldsymbol{y}^n$) is the sequence of n-grams of $\boldsymbol{x}$ ($\boldsymbol{y}$) (e.g., $\boldsymbol{x}^1 = \boldsymbol{x}$ is the sequence of words and $\boldsymbol{x}^2$ is the sequence of bigrams).

MoverScore-Main Idea

- The semantic distance is computed based on the Word Mover’s Distance (WMD).

$$\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \langle \boldsymbol{C}, \boldsymbol{F} \rangle,$$

$$\text{s.t. } \boldsymbol{F}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^{\intercal}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{y}^n}.$$

- System prediction $\boldsymbol{x} = (x_1, \ldots, x_m)$ is a sentence viewed as a sequence of words. Reference $\boldsymbol{y}$ is also a word sequence.

- $\boldsymbol{x}^n$ ($\boldsymbol{y}^n$) is the sequence of n-grams of $\boldsymbol{x}$ ($\boldsymbol{y}$) (e.g., $\boldsymbol{x}^1 = \boldsymbol{x}$ is the sequence of words and $\boldsymbol{x}^2$ is the sequence of bigrams).

- $\boldsymbol{C}$ is the transportation cost matrix such that $\boldsymbol{C}_{ij} = d(\boldsymbol{x}_i^n, \boldsymbol{y}_j^n)$ is the distance between the $\boldsymbol{i}$-th n-gram of $\boldsymbol{x}$ and the $\boldsymbol{j}$-th n-gram of $\boldsymbol{y}$.

MoverScore-Main Idea

- The semantic distance is computed based on the Word Mover's Distance (WMD).

$$\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \langle \boldsymbol{C}, \boldsymbol{F} \rangle,$$

$$\text{s.t. } \boldsymbol{F}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^{\intercal}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{y}^n}.$$

- System prediction $\boldsymbol{x} = (x_1, \ldots, x_m)$ is a sentence viewed as a sequence of words. Reference $\boldsymbol{y}$ is also a word sequence.

- $\boldsymbol{x}^n$ ($\boldsymbol{y}^n$) is the sequence of n-grams of $\boldsymbol{x}$ ($\boldsymbol{y}$) (e.g., $\boldsymbol{x}^1 = \boldsymbol{x}$ is the sequence of words and $\boldsymbol{x}^2$ is the sequence of bigrams).

- $\boldsymbol{C}$ is the transportation cost matrix such that $\boldsymbol{C}_{ij} = d(\boldsymbol{x}_i^n, \boldsymbol{y}_j^n)$ is the distance between the $\boldsymbol{i}$-th n-gram of $\boldsymbol{x}$ and the $\boldsymbol{j}$-th n-gram of $\boldsymbol{y}$.

- $\boldsymbol{f}_{\boldsymbol{x}^n} \in R_+^{|\boldsymbol{x}^n|}$ is a vector of weights. One weight for each n-gram of $\boldsymbol{x}^n$. We can assume $\boldsymbol{f}_{\boldsymbol{x}^n}^T \mathbf{1} = 1$, making $\boldsymbol{f}_{\boldsymbol{x}^n}$ a distribution over n-grams.

MoverScore-Main Idea

- The semantic distance is computed based on the Word Mover's Distance (WMD).

$$\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \langle \boldsymbol{C}, \boldsymbol{F} \rangle,$$

$$\text{s.t. } \boldsymbol{F}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^{\intercal}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{y}^n}.$$

- $\boldsymbol{F}$ is the transportation flow matrix with $\boldsymbol{F}_{ij}$ denoting the amount of flow traveling from the i-th n-gram x_i^n in $\boldsymbol{x}^n$ to the j-th n-gram y_j^n in $\boldsymbol{y}^n$.

- System prediction $\boldsymbol{x} = (x_1, \dots, x_m)$ is a sentence viewed as a sequence of words. Reference $\boldsymbol{y}$ is also a word sequence.

- $\boldsymbol{x}^n$ ($\boldsymbol{y}^n$) is the sequence of n-grams of $\boldsymbol{x}$ ($\boldsymbol{y}$) (e.g., $\boldsymbol{x}^1 = \boldsymbol{x}$ is the sequence of words and $\boldsymbol{x}^2$ is the sequence of bigrams).

- $\boldsymbol{C}$ is the transportation cost matrix such that $\boldsymbol{C}_{ij} = d(\boldsymbol{x}_i^n, \boldsymbol{y}_j^n)$ is the distance between the i-th n-gram of $\boldsymbol{x}$ and the j-th n-gram of $\boldsymbol{y}$.

- $\boldsymbol{f}_{\boldsymbol{x}^n} \in R_+^{|\boldsymbol{x}^n|}$ is a vector of weights. One weight for each n-gram of $\boldsymbol{x}^n$. We can assume $\boldsymbol{f}_{\boldsymbol{x}^n}^T \mathbf{1} = 1$, making $\boldsymbol{f}_{\boldsymbol{x}^n}$ a distribution over n-grams.

MoverScore-Main Idea

- The semantic distance is computed based on the Word Mover's Distance (WMD).

$$\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \langle \boldsymbol{C}, \boldsymbol{F} \rangle,$$

$$\text{s.t. } \boldsymbol{F}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^{\intercal}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{y}^n}.$$

- $\boldsymbol{F}$ is the transportation flow matrix with $\boldsymbol{F}_{ij}$ denoting the amount of flow traveling from the $\boldsymbol{i}$-th n-gram x_i^n in $\boldsymbol{x}^n$ to the $\boldsymbol{j}$-th n-gram y_j^n in $\boldsymbol{y}^n$.

- $\langle \boldsymbol{C}, \boldsymbol{F} \rangle$ denotes the sum of all matrix entries of the matrix $\boldsymbol{C} \odot \boldsymbol{F}$, where $\odot$ denotes element-wise multiplication.

- System prediction $\boldsymbol{x} = (x_1, \dots, x_m)$ is a sentence viewed as a sequence of words. Reference $\boldsymbol{y}$ is also a word sequence.

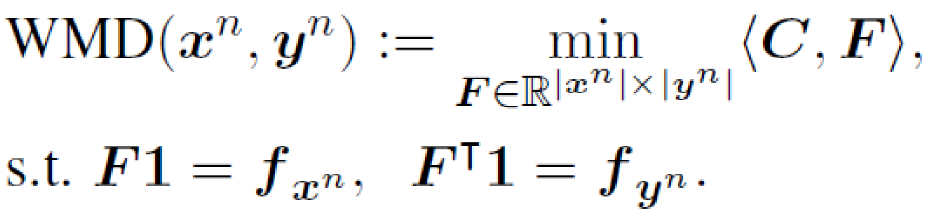
- $\boldsymbol{x}^n$ ($\boldsymbol{y}^n$) is the sequence of n-grams of $\boldsymbol{x}$ ($\boldsymbol{y}$) (e.g., $\boldsymbol{x}^1 = \boldsymbol{x}$ is the sequence of words and $\boldsymbol{x}^2$ is the sequence of bigrams).

- $\boldsymbol{C}$ is the transportation cost matrix such that $\boldsymbol{C}_{ij} = d(\boldsymbol{x}_i^n, \boldsymbol{y}_j^n)$ is the distance between the $\boldsymbol{i}$-th n-gram of $\boldsymbol{x}$ and the $\boldsymbol{j}$-th n-gram of $\boldsymbol{y}$.

- $\boldsymbol{f}_{\boldsymbol{x}^n} \in R_+^{|\boldsymbol{x}^n|}$ is a vector of weights. One weight for each n-gram of $\boldsymbol{x}^n$. We can assume $\boldsymbol{f}_{\boldsymbol{x}^n}^T \mathbf{1} = 1$, making $\boldsymbol{f}_{\boldsymbol{x}^n}$ a distribution over n-grams.

MoverScore-Main Idea

- The semantic distance is computed based on the Word Mover's Distance (WMD).



$$\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \langle \boldsymbol{C}, \boldsymbol{F} \rangle,$$

$$\text{s.t. } \boldsymbol{F}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^{\intercal}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{y}^n}.$$

- $\boldsymbol{F}$ is the transportation flow matrix with $\boldsymbol{F}_{ij}$ denoting the amount of flow traveling from the i-th n-gram x_i^n in $\boldsymbol{x}^n$ to the j-th n-gram y_j^n in $\boldsymbol{y}^n$.

- $\langle \boldsymbol{C}, \boldsymbol{F} \rangle$ denotes the sum of all matrix entries of the matrix $\boldsymbol{C} \odot \boldsymbol{F}$, where $\odot$ denotes element-wise multiplication.

- System prediction $\boldsymbol{x} = (x_1, \dots, x_m)$ is a sentence viewed as a sequence of words. Reference $\boldsymbol{y}$ is also a word sequence.

- $\boldsymbol{x}^n$ ($\boldsymbol{y}^n$) is the sequence of n-grams of $\boldsymbol{x}$ ($\boldsymbol{y}$) (e.g., $\boldsymbol{x}^1 = \boldsymbol{x}$ is the sequence of words and $\boldsymbol{x}^2$ is the sequence of bigrams).

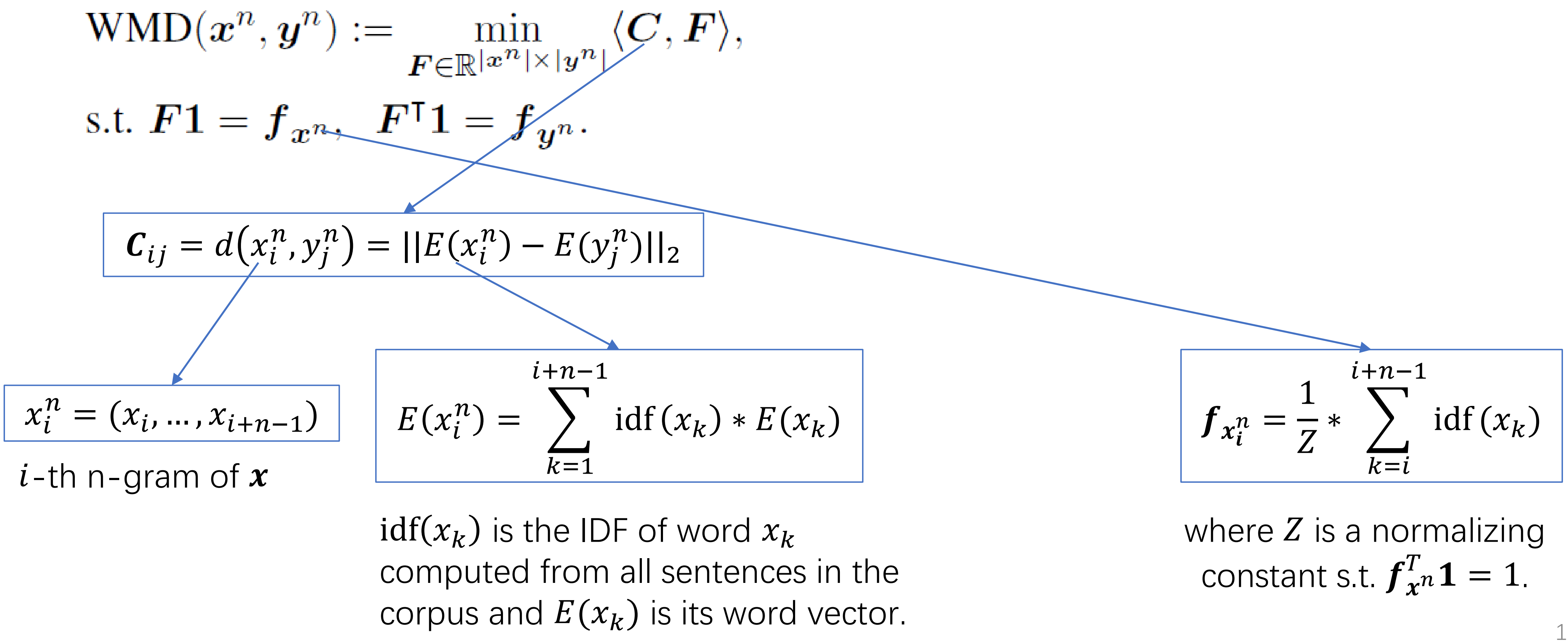
- $\boldsymbol{C}$ is the transportation cost matrix such that $\boldsymbol{C}_{ij} = d(\boldsymbol{x}_i^n, \boldsymbol{y}_j^n)$ is the distance between the i-th n-gram of $\boldsymbol{x}$ and the j-th n-gram of $\boldsymbol{y}$.

- $\boldsymbol{f}_{\boldsymbol{x}^n} \in R_+^{|\boldsymbol{x}^n|}$ is a vector of weights. One weight for each n-gram of $\boldsymbol{x}^n$. We can assume $\boldsymbol{f}_{\boldsymbol{x}^n}^T \mathbf{1} = 1$, making $\boldsymbol{f}_{\boldsymbol{x}^n}$ a distribution over n-grams.

Insight: find the minimum effort to transform between two texts

MoverScore-In Practice

- The semantic distance is computed based on the Word Mover's Distance (WMD).



$$\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F}\in\mathbb{R}^{|\boldsymbol{x}^n|\times|\boldsymbol{y}^n|}} \langle \boldsymbol{C}, \boldsymbol{F}\rangle,$$

$$\text{s.t. } \boldsymbol{F}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^\intercal\mathbf{1} = \boldsymbol{f}_{\boldsymbol{y}^n}.$$

$$\boldsymbol{C}_{ij} = d(x_i^n, y_j^n) = ||E(x_i^n) - E(y_j^n)||_2$$

$$x_i^n = (x_i, \dots, x_{i+n-1})$$

i-th n-gram of $\boldsymbol{x}$

$$E(x_i^n) = \sum_{k=1}^{i+n-1} \mathrm{idf}\,(x_k) * E(x_k)$$

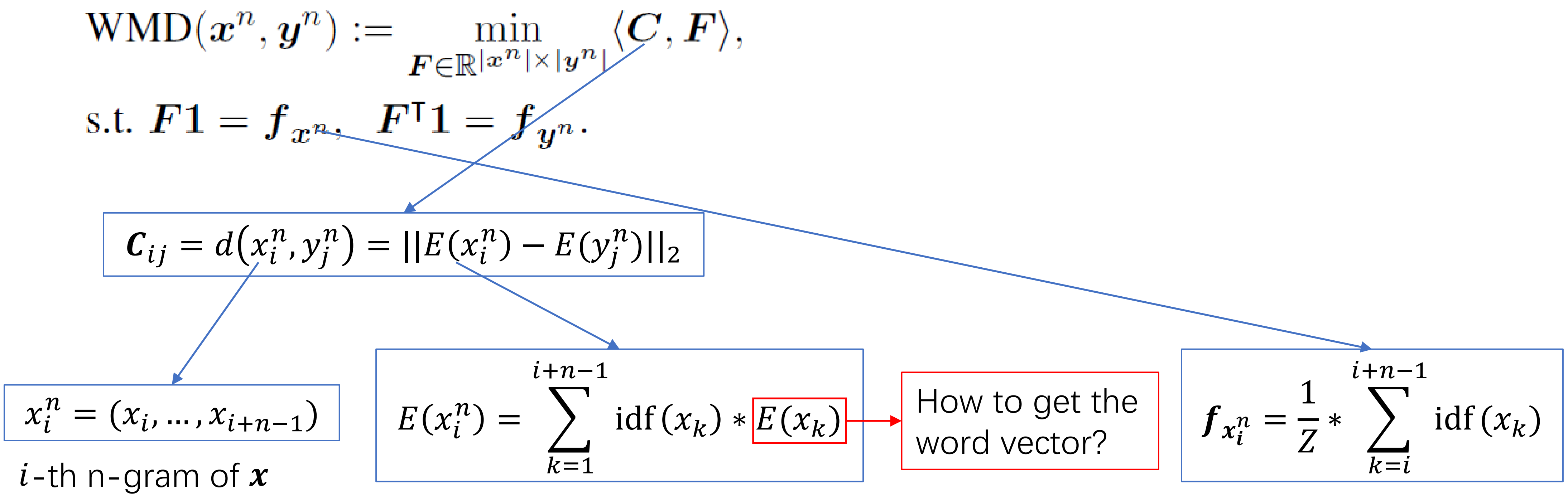
$\mathrm{idf}(x_k)$ is the IDF of word x_k computed from all sentences in the corpus and $E(x_k)$ is its word vector.

$$\boldsymbol{f}_{\boldsymbol{x}_i^n} = \frac{1}{Z} * \sum_{k=i}^{i+n-1} \mathrm{idf}\,(x_k)$$

where Z is a normalizing constant s.t. $\boldsymbol{f}_{\boldsymbol{x}^n}^T\mathbf{1} = 1$.

MoverScore-In Practice

- The semantic distance is computed based on the Word Mover's Distance (WMD).



$$\mathrm{WMD}(\boldsymbol{x}^n, \boldsymbol{y}^n) := \min_{\boldsymbol{F} \in \mathbb{R}^{|\boldsymbol{x}^n| \times |\boldsymbol{y}^n|}} \langle \boldsymbol{C}, \boldsymbol{F} \rangle,$$

$$\text{s.t. } \boldsymbol{F}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{x}^n}, \quad \boldsymbol{F}^{\intercal}\mathbf{1} = \boldsymbol{f}_{\boldsymbol{y}^n}.$$

$$\boldsymbol{C}_{ij} = d(x_i^n, y_j^n) = ||E(x_i^n) - E(y_j^n)||_2$$

$$x_i^n = (x_i, \ldots, x_{i+n-1})$$

i-th n-gram of $\boldsymbol{x}$

$$E(x_i^n) = \sum_{k=1}^{i+n-1} \mathrm{idf}(x_k) * \boxed{E(x_k)}$$

How to get the word vector?

$$\boldsymbol{f}_{\boldsymbol{x}_i^n} = \frac{1}{Z} * \sum_{k=i}^{i+n-1} \mathrm{idf}(x_k)$$

$\mathrm{idf}(x_k)$ is the IDF of word x_k computed from all sentences in the corpus and $E(x_k)$ is its word vector.

where Z is a normalizing constant s.t. $\boldsymbol{f}_{\boldsymbol{x}^n}^T \mathbf{1} = 1$.

MoverScore-In Practice

- How to get the word vector?

$E(x_i)$ {
Static embeddings, e.g. word2vec

Contextualized embeddings, e.g. ELMo, BERT

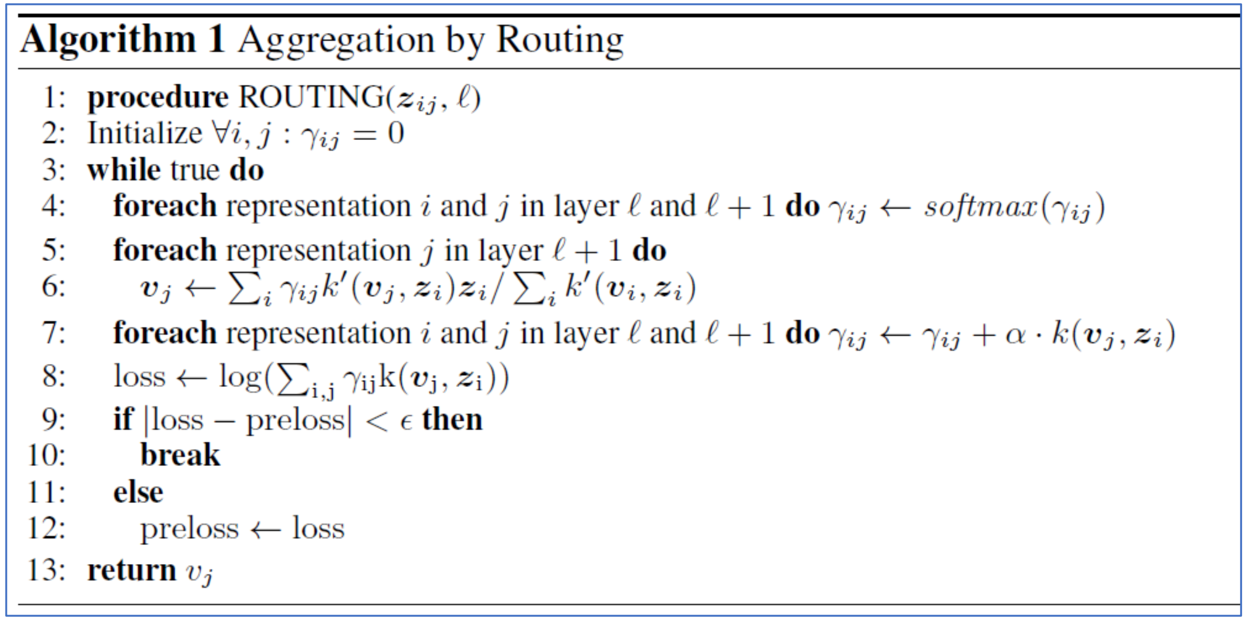
- If choose the contextualized embeddings, how to aggregate the word vectors from multiple (e.g. L) layers?

Power Means

$$E(x_i) = \mathbf{h}_i^{(1)} \oplus \mathbf{h}_i^{(+\infty)} \oplus \mathbf{h}_i^{(-\infty)}$$

$$\mathbf{h}_i^{(p)} = \left(\frac{\mathbf{z}_{i,1}^p + \cdots + \mathbf{z}_{i,L}^p}{L} \right)^{\frac{1}{p}}$$

Power Means



Algorithm 1 Aggregation by Routing

```
procedure ROUTING($z_{ij}, \ell$)
Initialize $\forall i, j : \gamma_{ij} = 0$
while true do
  foreach representation $i$ and $j$ in layer $\ell$ and $\ell + 1$ do $\gamma_{ij} \leftarrow softmax(\gamma_{ij})$
  foreach representation $j$ in layer $\ell + 1$ do
    $v_j \leftarrow \sum_i \gamma_{ij} k'(v_j, z_i) z_i / \sum_i k'(v_i, z_i)$
  foreach representation $i$ and $j$ in layer $\ell$ and $\ell + 1$ do $\gamma_{ij} \leftarrow \gamma_{ij} + \alpha \cdot k(v_j, z_i)$
  loss $\leftarrow \log(\sum_{i,j} \gamma_{ij} k(v_j, z_i))$
  if $|$loss $-$ preloss$| < \epsilon$ then
    break
  else
    preloss $\leftarrow$ loss
return $v_j$
```

Routing

MoverScore-In Practice

- Sentence Mover Distance (SMD) is computed from the distance between the two sentence embeddings.

$$\mathrm{SMD}(\boldsymbol{x}^{\mathrm{n}}, \boldsymbol{y}^{\mathrm{n}}) = ||E\left(x_1^{l_x}\right) - E\left(y_1^{l_y}\right)||_2$$

where l_x and l_y are the size of sentences

MoverScore-Experimental Setup

- The MoverScore has been investigated along four dimensions:

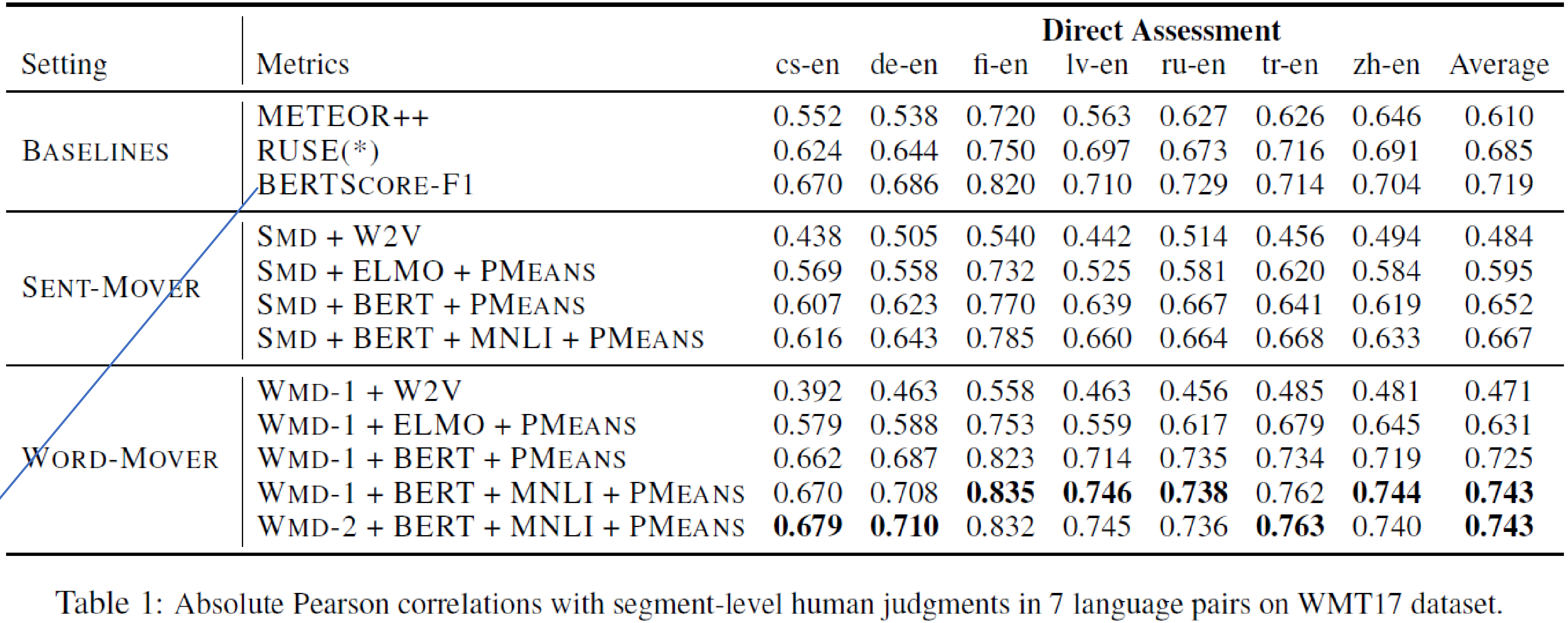
  1. the granularity of embeddings, i.e., the size of n for n-grams
     - n=1
     - n=2
     - n=sentence length
  2. the choice of pretrained embedding mechanism
     - word2vec
     - ELMo
     - BERT
  3. the fine-tuning task used for BERT
     - MultiNLI
     - QANLI
     - QQP
  4. the aggregation technique (p-means or routing) when applicable
     - p-means
     - Routing

e.g., WMD-1+BERT+MNLI+PMEANS

- The major focus is to study the correlation between different metrics and human judgment. Pearson's r and Spearman's ρ are selected to measure the correlation.

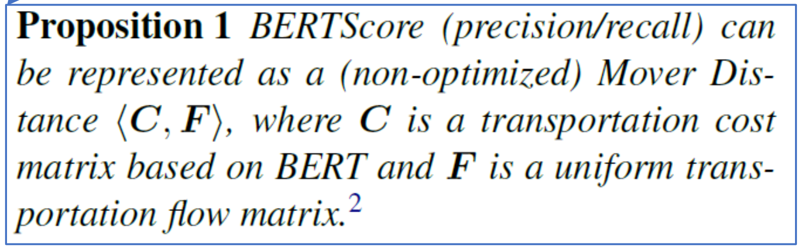
MoverScore-Experiments on Translation

- Dataset: WMT 2017; 7 language pairs; Each language pair has approximately 3,000 sentences.



Setting	Metrics	Direct Assessment							
		cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Average
BASELINES	METEOR++	0.552	0.538	0.720	0.563	0.627	0.626	0.646	0.610
	RUSE(*)	0.624	0.644	0.750	0.697	0.673	0.716	0.691	0.685
	BERTSCORE-F1	0.670	0.686	0.820	0.710	0.729	0.714	0.704	0.719
SENT-MOVER	SMD + W2V	0.438	0.505	0.540	0.442	0.514	0.456	0.494	0.484
	SMD + ELMO + PMEANS	0.569	0.558	0.732	0.525	0.581	0.620	0.584	0.595
	SMD + BERT + PMEANS	0.607	0.623	0.770	0.639	0.667	0.641	0.619	0.652
	SMD + BERT + MNLI + PMEANS	0.616	0.643	0.785	0.660	0.664	0.668	0.633	0.667
WORD-MOVER	WMD-1 + W2V	0.392	0.463	0.558	0.463	0.456	0.485	0.481	0.471
	WMD-1 + ELMO + PMEANS	0.579	0.588	0.753	0.559	0.617	0.679	0.645	0.631
	WMD-1 + BERT + PMEANS	0.662	0.687	0.823	0.714	0.735	0.734	0.719	0.725
	WMD-1 + BERT + MNLI + PMEANS	0.670	0.708	**0.835**	**0.746**	**0.738**	0.762	**0.744**	**0.743**
	WMD-2 + BERT + MNLI + PMEANS	**0.679**	**0.710**	0.832	0.745	0.736	**0.763**	0.740	**0.743**

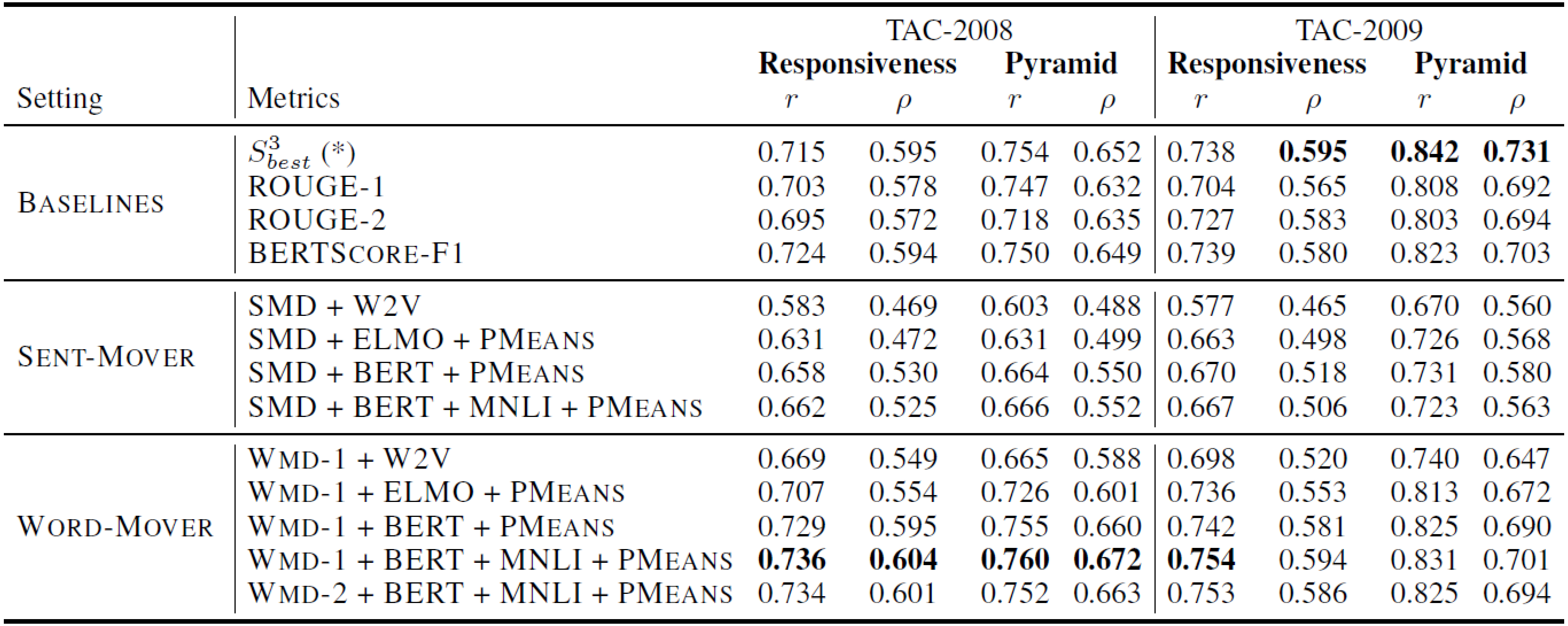
Table 1: Absolute Pearson correlations with segment-level human judgments in 7 language pairs on WMT17 dataset.



Proposition 1 *BERTScore (precision/recall) can be represented as a (non-optimized) Mover Distance $\langle \boldsymbol{C}, \boldsymbol{F} \rangle$, where $\boldsymbol{C}$ is a transportation cost matrix based on BERT and $\boldsymbol{F}$ is a uniform transportation flow matrix.*[2]

MoverScore-Experiments on Summarization

- Datasets: TAC2008/TAC2009; 48/44 clusters; 10 news article per cluster; four reference summaries per cluster;

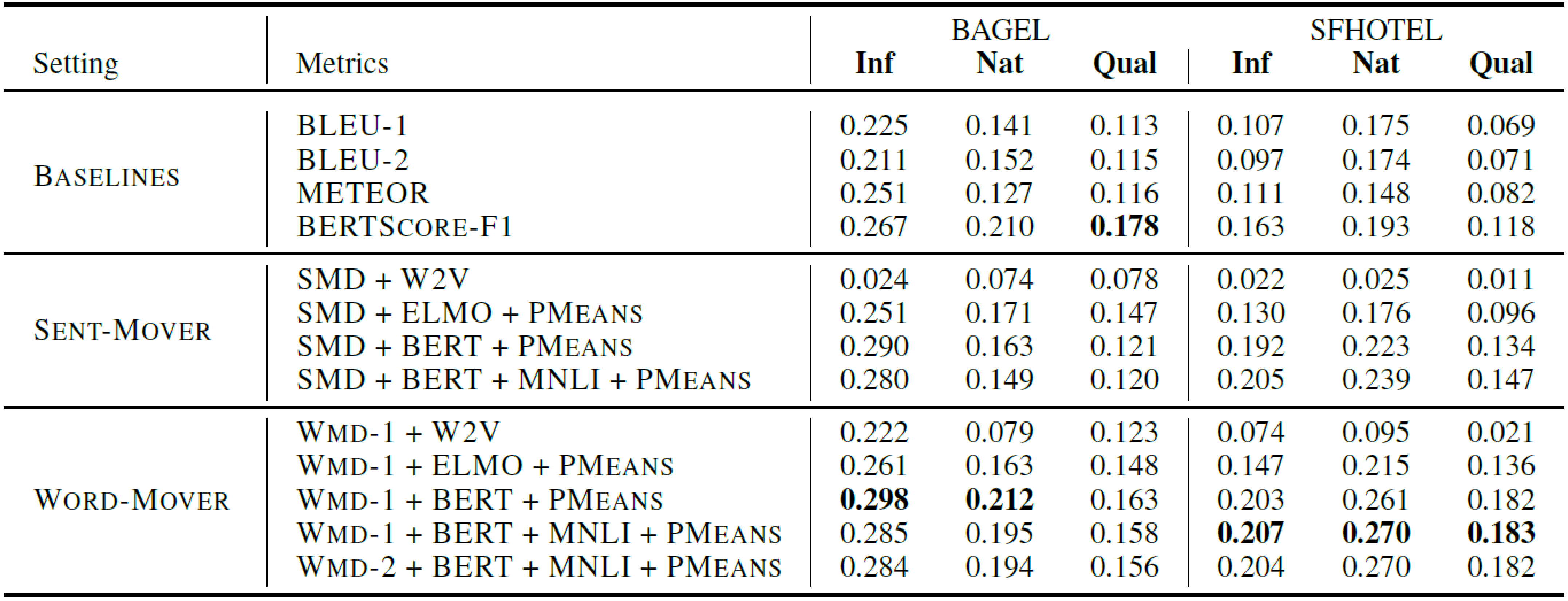


		TAC-2008				TAC-2009			
		Responsiveness		**Pyramid**		**Responsiveness**		**Pyramid**	
Setting	Metrics	r	ρ	r	ρ	r	ρ	r	ρ
BASELINES	S^3_{best} (*)	0.715	0.595	0.754	0.652	0.738	**0.595**	**0.842**	**0.731**
	ROUGE-1	0.703	0.578	0.747	0.632	0.704	0.565	0.808	0.692
	ROUGE-2	0.695	0.572	0.718	0.635	0.727	0.583	0.803	0.694
	BERTSCORE-F1	0.724	0.594	0.750	0.649	0.739	0.580	0.823	0.703
SENT-MOVER	SMD + W2V	0.583	0.469	0.603	0.488	0.577	0.465	0.670	0.560
	SMD + ELMO + PMEANS	0.631	0.472	0.631	0.499	0.663	0.498	0.726	0.568
	SMD + BERT + PMEANS	0.658	0.530	0.664	0.550	0.670	0.518	0.731	0.580
	SMD + BERT + MNLI + PMEANS	0.662	0.525	0.666	0.552	0.667	0.506	0.723	0.563
WORD-MOVER	WMD-1 + W2V	0.669	0.549	0.665	0.588	0.698	0.520	0.740	0.647
	WMD-1 + ELMO + PMEANS	0.707	0.554	0.726	0.601	0.736	0.553	0.813	0.672
	WMD-1 + BERT + PMEANS	0.729	0.595	0.755	0.660	0.742	0.581	0.825	0.690
	WMD-1 + BERT + MNLI + PMEANS	**0.736**	**0.604**	**0.760**	**0.672**	**0.754**	0.594	0.831	0.701
	WMD-2 + BERT + MNLI + PMEANS	0.734	0.601	0.752	0.663	0.753	0.586	0.825	0.694

Table 2: Pearson r and Spearman ρ correlations with summary-level human judgments on TAC 2008 and 2009.

MoverScore-Experiments on Dialogue

- Datasets: BAGEL/SFHOTEL; 202/398 instances with multiple references;

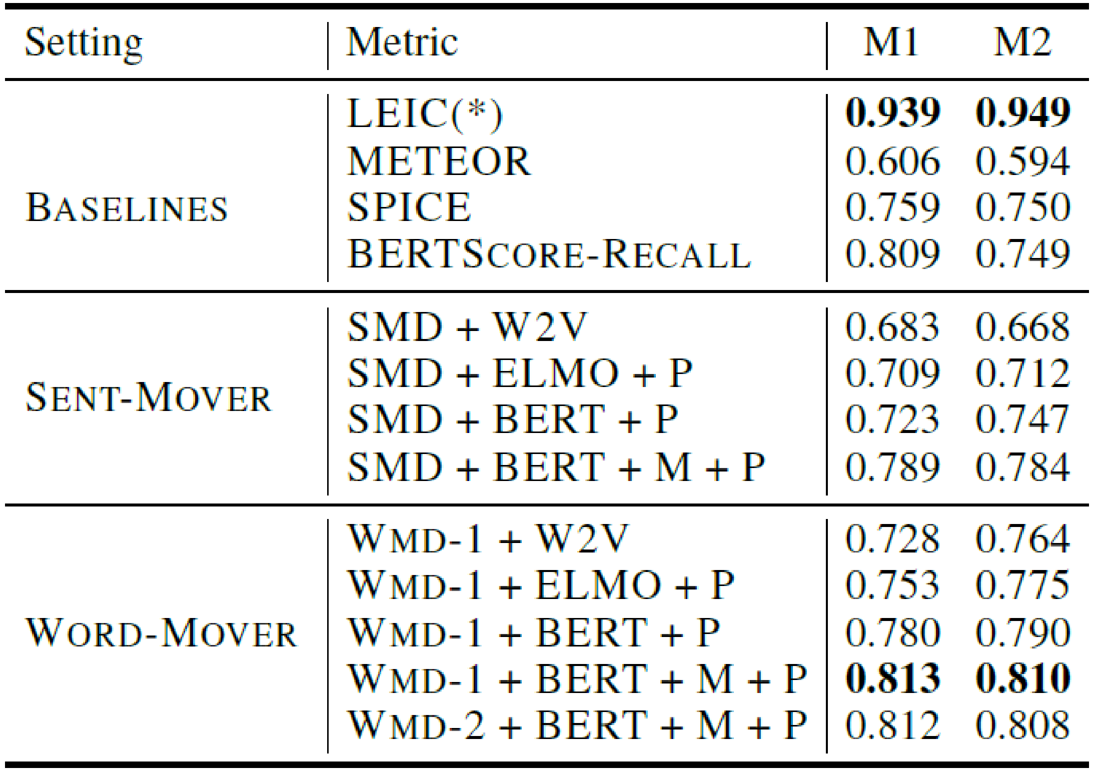


			BAGEL			SFHOTEL	
Setting	Metrics	**Inf**	**Nat**	**Qual**	**Inf**	**Nat**	**Qual**
Baselines	BLEU-1	0.225	0.141	0.113	0.107	0.175	0.069
	BLEU-2	0.211	0.152	0.115	0.097	0.174	0.071
	METEOR	0.251	0.127	0.116	0.111	0.148	0.082
	BERTScore-F1	0.267	0.210	**0.178**	0.163	0.193	0.118
Sent-Mover	SMD + W2V	0.024	0.074	0.078	0.022	0.025	0.011
	SMD + ELMO + PMeans	0.251	0.171	0.147	0.130	0.176	0.096
	SMD + BERT + PMeans	0.290	0.163	0.121	0.192	0.223	0.134
	SMD + BERT + MNLI + PMeans	0.280	0.149	0.120	0.205	0.239	0.147
Word-Mover	WMD-1 + W2V	0.222	0.079	0.123	0.074	0.095	0.021
	WMD-1 + ELMO + PMeans	0.261	0.163	0.148	0.147	0.215	0.136
	WMD-1 + BERT + PMeans	**0.298**	**0.212**	0.163	0.203	0.261	0.182
	WMD-1 + BERT + MNLI + PMeans	0.285	0.195	0.158	**0.207**	**0.270**	**0.183**
	WMD-2 + BERT + MNLI + PMeans	0.284	0.194	0.156	0.204	0.270	0.182

Table 3: Spearman correlation with utterance-level human judgments for BAGEL and SFHOTEL datasets.

MoverScore-Experiments on Image Caption

- Dataset: MSCOCO; 5000 instances; five caption references per instance;



Setting	Metric	M1	M2
BASELINES	LEIC(*)	**0.939**	**0.949**
	METEOR	0.606	0.594
	SPICE	0.759	0.750
	BERTSCORE-RECALL	0.809	0.749
SENT-MOVER	SMD + W2V	0.683	0.668
	SMD + ELMO + P	0.709	0.712
	SMD + BERT + P	0.723	0.747
	SMD + BERT + M + P	0.789	0.784
WORD-MOVER	WMD-1 + W2V	0.728	0.764
	WMD-1 + ELMO + P	0.753	0.775
	WMD-1 + BERT + P	0.780	0.790
	WMD-1 + BERT + M + P	**0.813**	**0.810**
	WMD-2 + BERT + M + P	0.812	0.808

Table 4: Pearson correlation with system-level human judgments on MSCOCO dataset. 'M' and 'P' are short names.

MoverScore-Experiments

- Score distribution

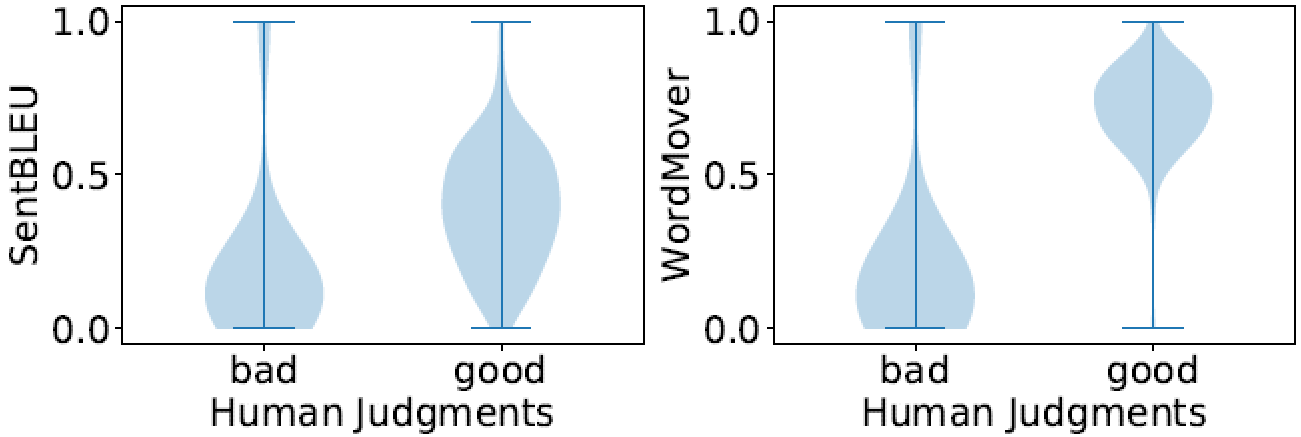


Figure 2: Score distribution in German-to-English pair.

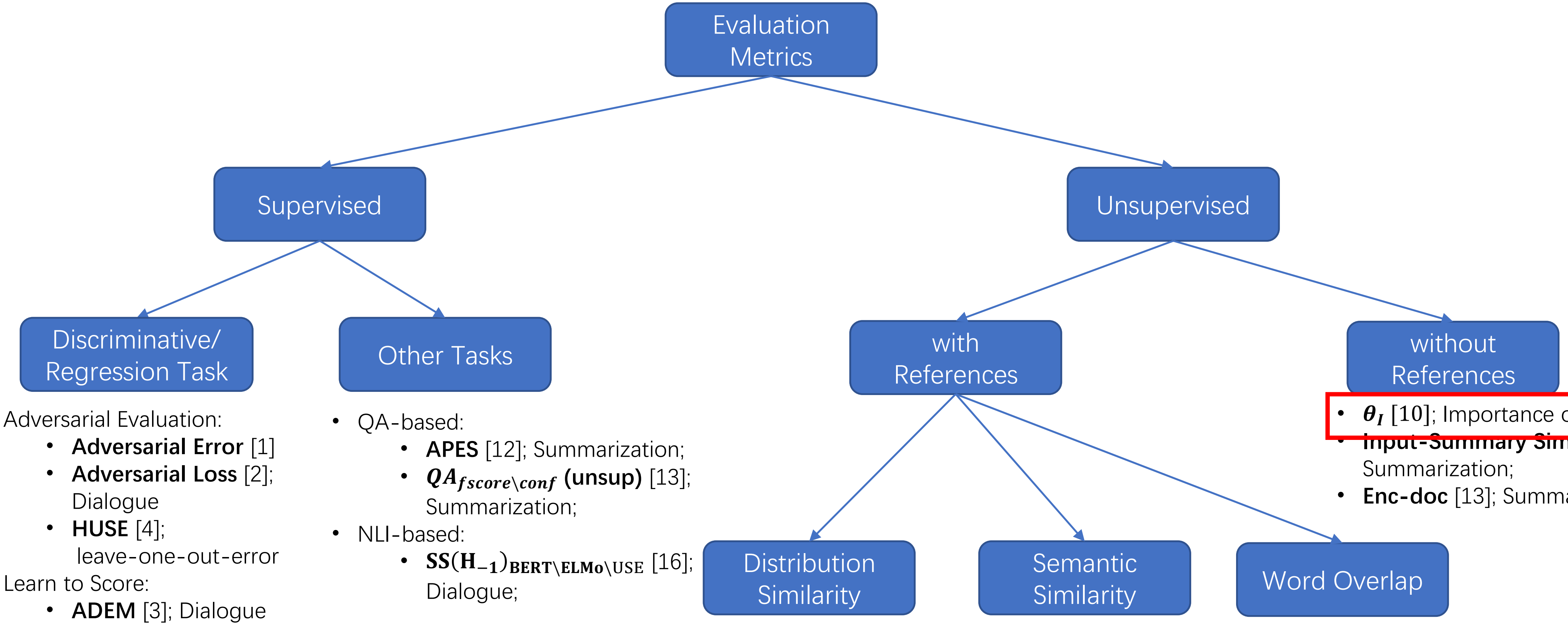
MoverScore-Conclusions

- Investigated new unsupervised evaluation metrics for text generation systems combining contextualized embeddings with Earth Mover’s Distance.

- The new metric obtain strong generalization ability across four text generation tasks, oftentimes even outperforming supervised metrics.

- One limitation of this metric is that it depends on the IDF of generated summaries. When adding a new system to evaluate, the scores of other systems will be changed.
  - BERTSCORE has no such limitation.

Brief Taxonomy



- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

- QA-based:
  - **APES** [12]; Summarization;
  - $\boldsymbol{QA_{fscore\backslash conf}}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

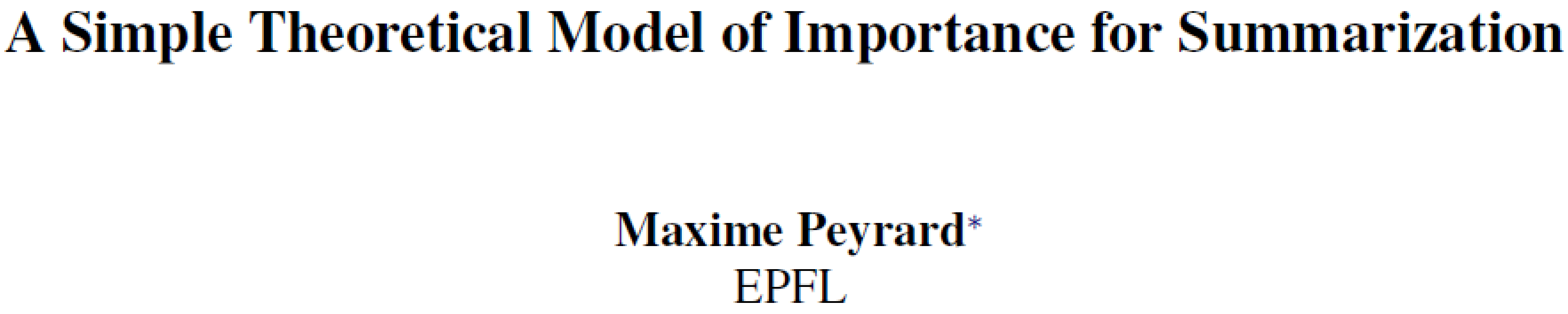
- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

- **Fre´chet BERT Distance (FBD)** [5]

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

- **ROUGE; BLEU; ⋯**
- **PyrEval** [11]; Summarization;

θ_I - Title & Authors



A Simple Theoretical Model of Importance for Summarization

Maxime Peyrard*
EPFL

θ_I -Introduction

- **Motivation:** the notion of information Importance remains latent in summarization research.

- **Method:** propose simple theoretical models of Importance by unifying the following concepts:
  - Redundancy
  - Relevance
  - Informativeness

- **Contributions:**
  1. define several concepts intuitively connected to summarization: *Redundancy*, *Relevance* and *Informativeness*.
  2. formulate properties required from a useful notion of *Importance* as the quantity unifying these concepts & provide intuitions to interpret the proposed quantities.
  3. even under simplifying assumptions, these quantities correlates well with human judgments

θ_I-Redundancy

- In information-theoretic terms, the amount of information is measured by Shannon's entropy. For a summary S represented by P_S:

$$H(S) = -\sum_{w_i} P_S(w_i)\log\left(P_S(w_i)\right)$$

semantic unit
e.g., word

e.g., word frequency distribution

θ_I-Redundancy

- In information-theoretic terms, the amount of information is measured by Shannon's entropy. For a summary S represented by P_S:

$$H(S) = -\sum_{w_i} P_S(w_i)\log(P_S(w_i))$$

semantic unit
e.g., word

e.g., word frequency distribution

- The *Redundancy* is defined as:

$$Red(S) = H_{max} - H(S)$$

θ_I-Redundancy

- In information-theoretic terms, the amount of information is measured by Shannon's entropy. For a summary S represented by P_S:

$$H(S) = -\sum_{w_i} P_S(w_i)\log\left(P_S(w_i)\right)$$

semantic unit
e.g., word

e.g., word frequency distribution

- The *Redundancy* is defined as:

$$Red(S) = H_{max} - H(S)$$

$$\Big\downarrow \; H_{max} \text{ is a constant}$$

$$Red(S) = -H(S)$$

θ_I -Relevance

- estimating *Relevance* boils down to comparing the distributions P_S and P_D (*D* is the document), which is done via the cross-entropy:

$$Rel(S, D) = -CE(S, D) = \sum_{w_i} P_S(w_i) \log\left(P_D(w_i)\right)$$

The cross-entropy is interpreted as the average surprise of observing S while expecting D. Lower surprise indicates higher relevance.

- $-KL(S||D) = Rel(S, D) - Red(S)$

Maximizing *Relevance* & Minimizing *Redundancy* = Minimizing the KL divergence between P_S and P_D

θ_I -Informativeness

- Intuitively, a summary is informative if it induces, for a user, a great change in her/his knowledge about the world.
- We denote the background knowledge as K which is represented by a probability distribution P_K over semantic units.
- *Informativeness* is defined as the amount of new information contained in a summary S compared to K. It can be given by the cross entropy:

$$Inf(S,K) = CE(S,K) = -\sum_{w_i} P_S(w_i)\log(P_K(w_i))$$

The cross-entropy is interpreted as the average surprise of observing S while expecting K. Higher surprise indicates higher *Informativeness*.

$\boldsymbol{\theta_I}$ -The Unified Importance

$$\theta_I(S,D,K) \equiv -Red(S) + \alpha * Rel(S,D) + \beta * Inf(S,K)$$

$$Red(S) = -H(S)$$

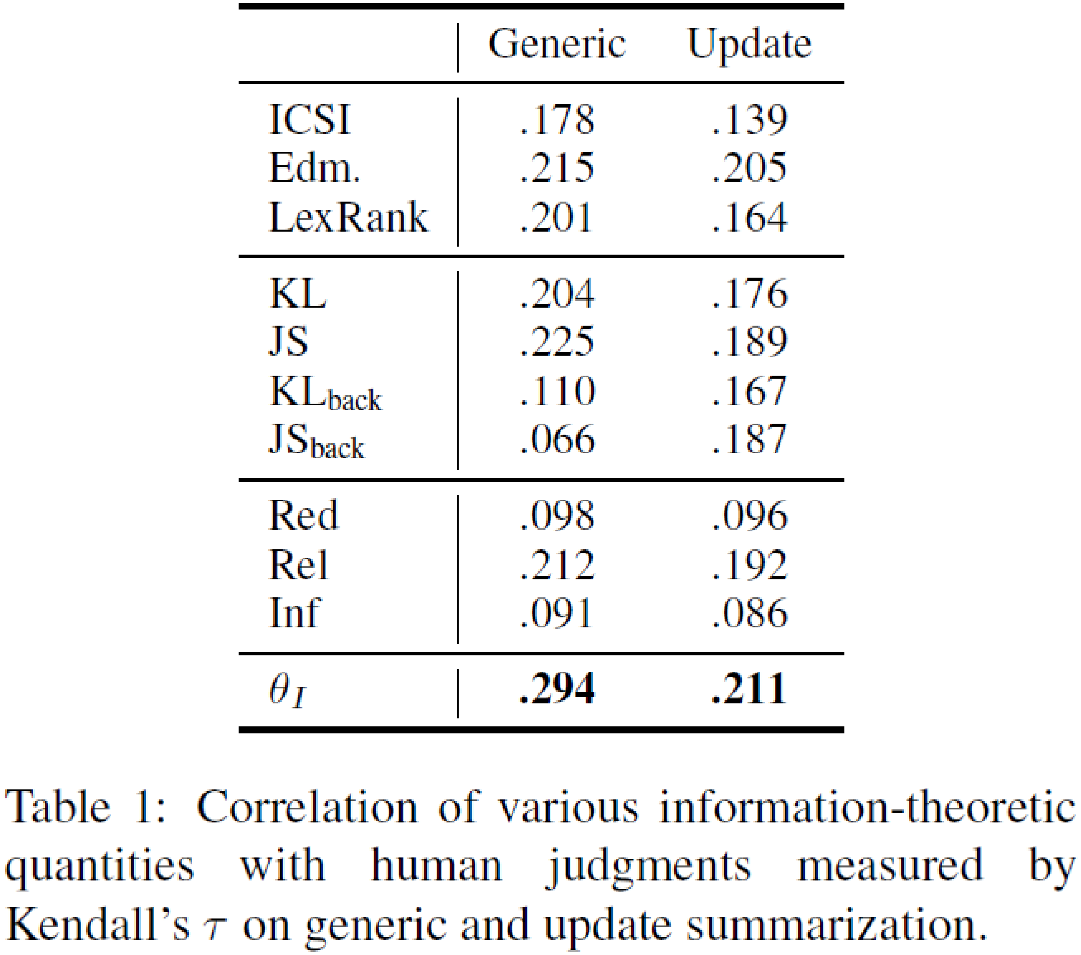
$$Rel(S,D) = -CE(S,D) = \sum_{w_i} P_S(w_i)\log\left(P_D(w_i)\right)$$

$$Inf(S,K) = CE(S,K) = -\sum_{w_i} P_S(w_i)\log\left(P_K(w_i)\right)$$

θ_I - Experiments

- Choose word as the semantic unit.
- Texts are represented frequency distribution over words.
- $\alpha = \beta = 1$
- Datasets: TAC-2008; TAC-2009;
- Two summarization settings:
  - Generic multi-document summarization
    - 10 documents (A documents) are to be summarized.
    - K is the uniform probability distribution over all words from the source documents.
  - Update multi-document summarization
    - 10 new documents (B documents) are to be summarized assuming that the first 10 documents (A documents) have already been seen.
    - K is the frequency distribution over words in the background documents (A).

θ_I-Experiments



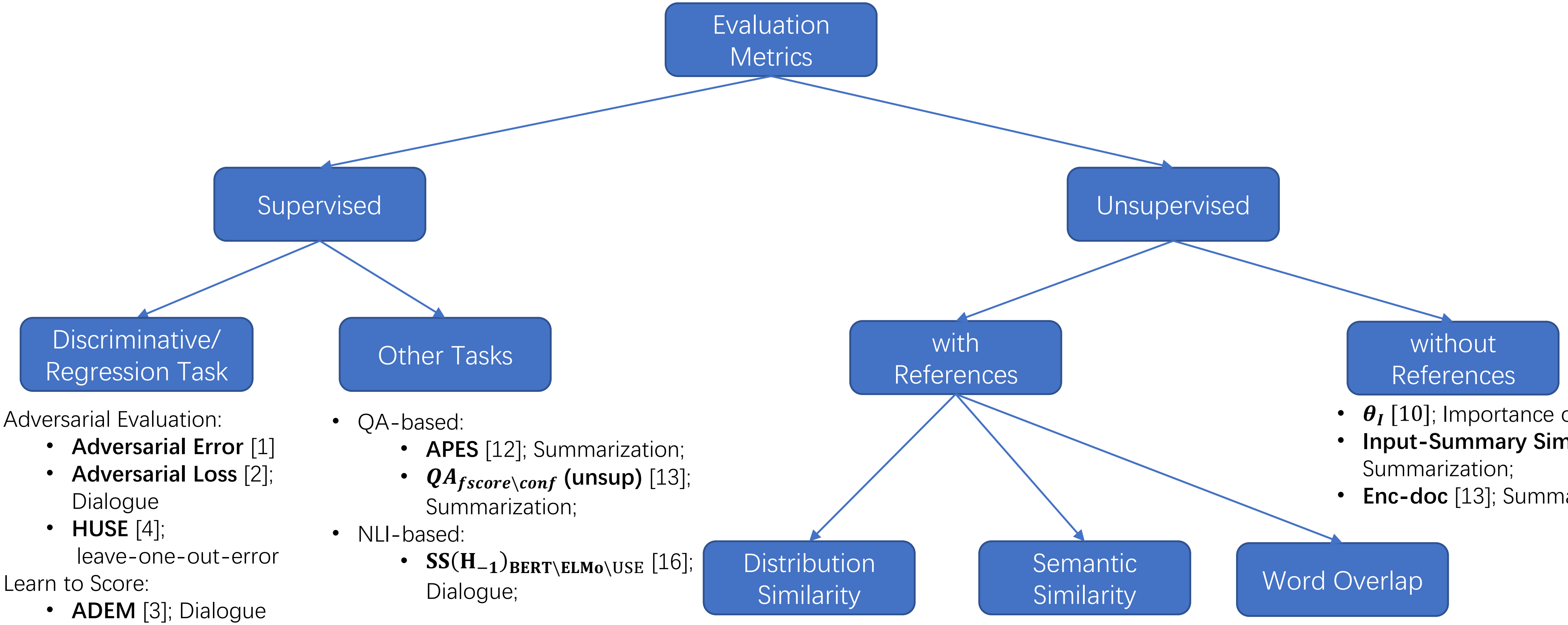
	Generic	Update
ICSI	.178	.139
Edm.	.215	.205
LexRank	.201	.164
KL	.204	.176
JS	.225	.189
KL_{back}	.110	.167
JS_{back}	.066	.187
Red	.098	.096
Rel	.212	.192
Inf	.091	.086
θ_I	**.294**	**.211**

Table 1: Correlation of various information-theoretic quantities with human judgments measured by Kendall's τ on generic and update summarization.

θ_I - Conclusions

- A simple theoretical modeling of summary *Importance* with elegant and self-contained interpretation.
- Generalization ability is not good enough since it seems to be specifically-designed for multi-document summarization.

Brief Taxonomy



Evaluation Metrics

Supervised

Unsupervised

Discriminative/ Regression Task

- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

Other Tasks

- QA-based:
  - **APES** [12]; Summarization;
  - $\boldsymbol{QA_{fscore\backslash conf}}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

with References

Distribution Similarity

- **Fre´chet BERT Distance (FBD)** [5]

Semantic Similarity

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

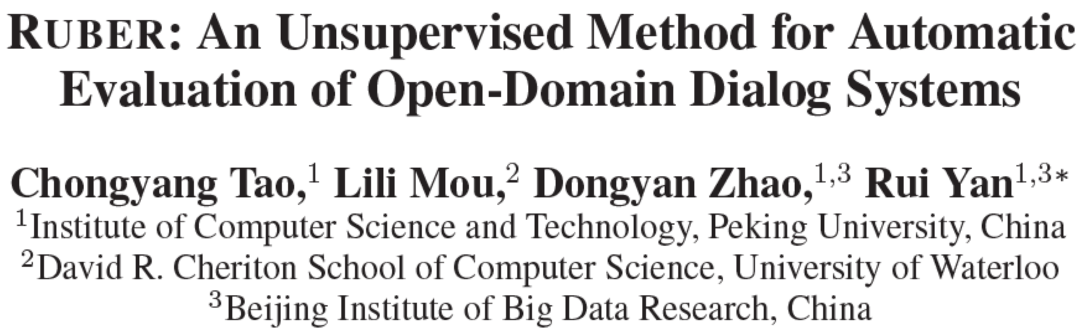
Word Overlap

- **ROUGE; BLEU; ⋯**
- **PyrEval** [11]; Summarization;

without References

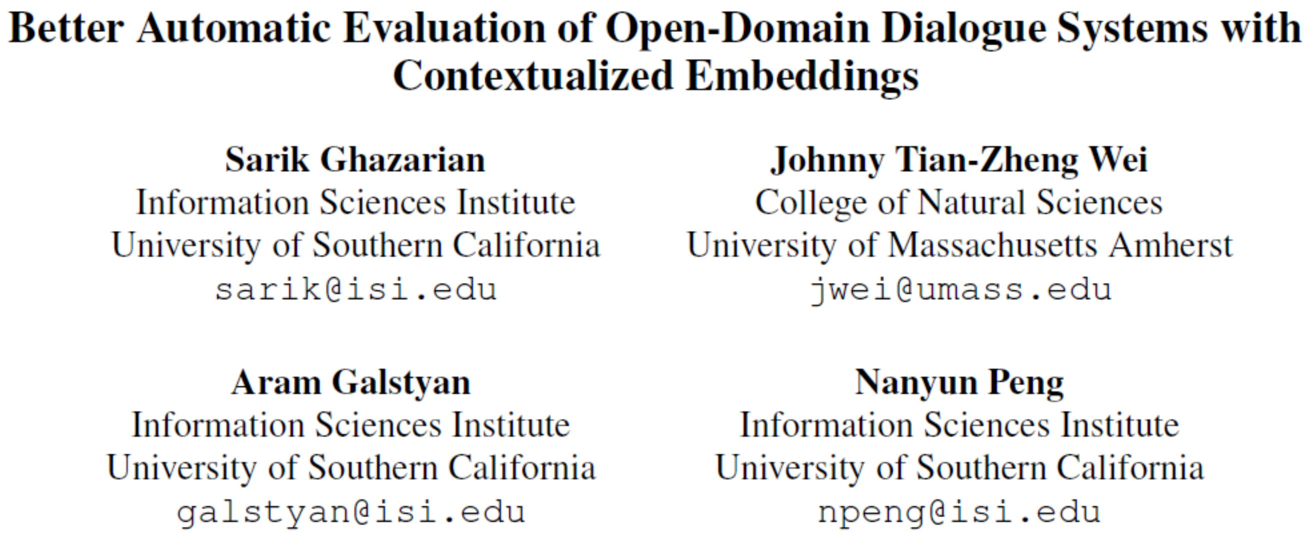
- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

RUBER-Title & Authors



RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems

Chongyang Tao,[1] **Lili Mou,**[2] **Dongyan Zhao,**[1,3] **Rui Yan**[1,3]*
[1]Institute of Computer Science and Technology, Peking University, China
[2]David R. Cheriton School of Computer Science, University of Waterloo
[3]Beijing Institute of Big Data Research, China



Better Automatic Evaluation of Open-Domain Dialogue Systems with Contextualized Embeddings

Sarik Ghazarian
Information Sciences Institute
University of Southern California
sarik@isi.edu

Johnny Tian-Zheng Wei
College of Natural Sciences
University of Massachusetts Amherst
jwei@umass.edu

Aram Galstyan
Information Sciences Institute
University of Southern California
galstyan@isi.edu

Nanyun Peng
Information Sciences Institute
University of Southern California
npeng@isi.edu

RUBER-Introduction

- **Motivation:** researchers usually resort to human annotation for dialogue model evaluation, which is time and labor-intensive.

- **Method:** blend a referenced metric and unreferenced metric as the final metric.

- **Contributions:**
  1. Referenced metric. An embedding-based scorer measures the similarity between a generated reply and the ground truth.
  2. Unreferenced metric. A neural network-based scorer measures the relatedness between the generated reply and its query.
  3. RUBER. Combining the referenced and unreferenced metrics to better make use of both worlds.

RUBER-Methodology

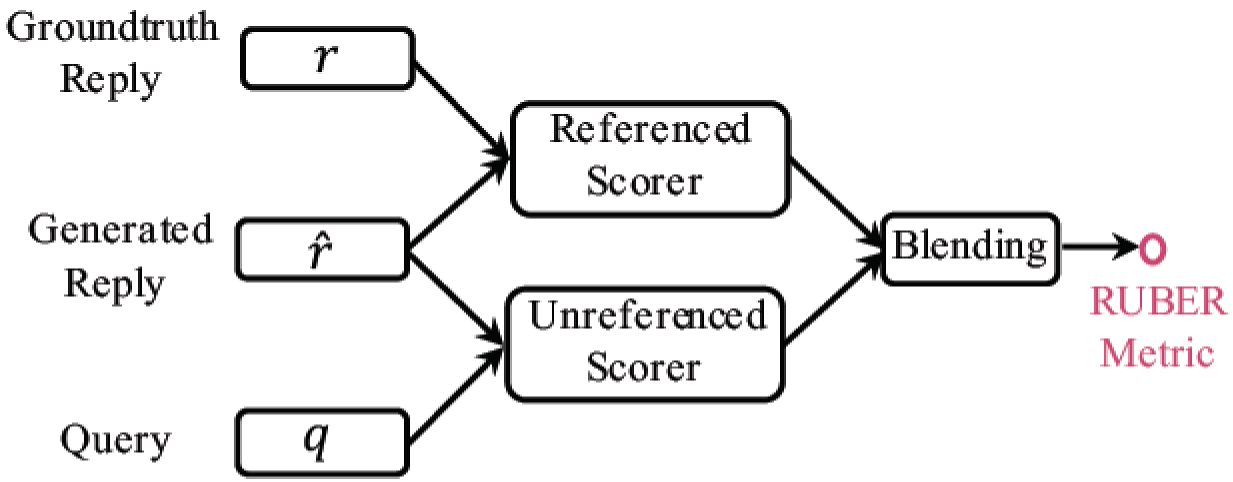
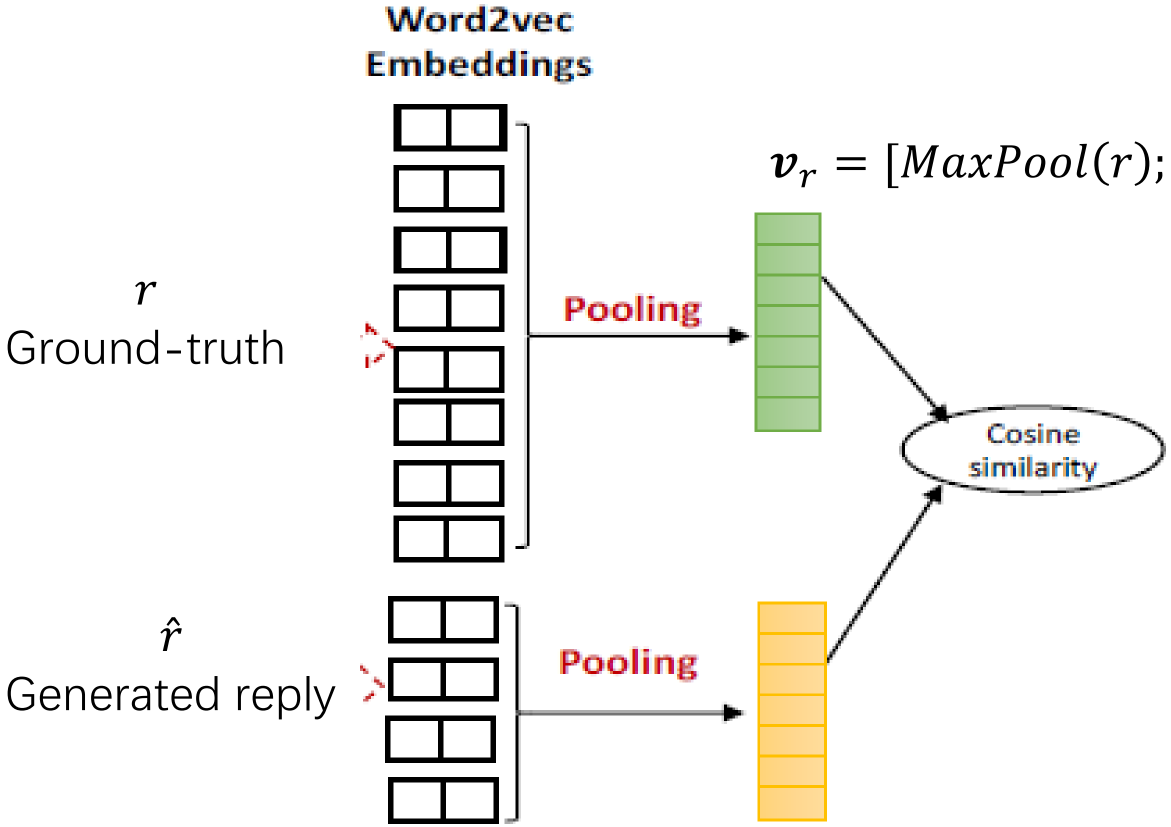


Figure 2: Overview of the RUBER metric.

RUBER-Methodology

- Referenced Metric



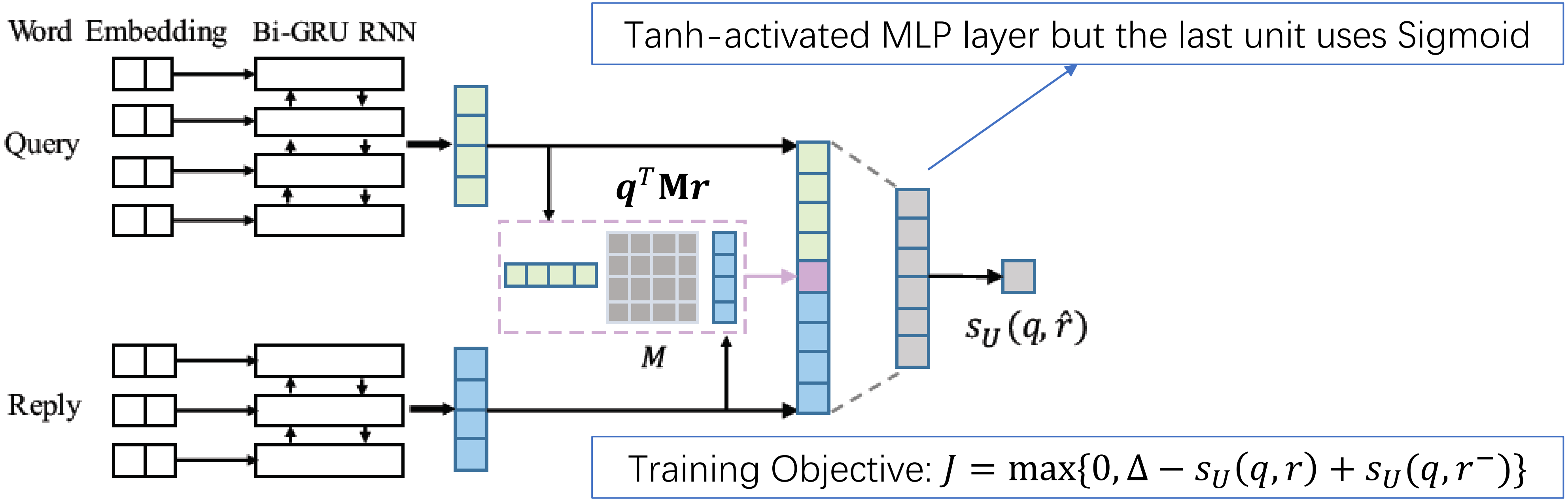
$$\boldsymbol{v}_r = [MaxPool(r); MinPool(r)]$$

$$s_R(r, \hat{r}) = \cos(\boldsymbol{v}_r, \boldsymbol{v}_{\hat{r}}) = \frac{\boldsymbol{v}_r^T \boldsymbol{v}_{\hat{r}}}{||\boldsymbol{v}_r|| * ||\boldsymbol{v}_{\hat{r}}||}$$

$$\boldsymbol{v}_{\hat{r}} = [MaxPool(\hat{r}); MinPool(\hat{r})]$$

RUBER-Methodology

- Unreferenced Metric



Tanh-activated MLP layer but the last unit uses Sigmoid

Training Objective: $J = \max\{0, \Delta - s_U(q,r) + s_U(q,r^-)\}$

Figure 3: The neural network predicting the unreferenced score.

RUBER-Methodology

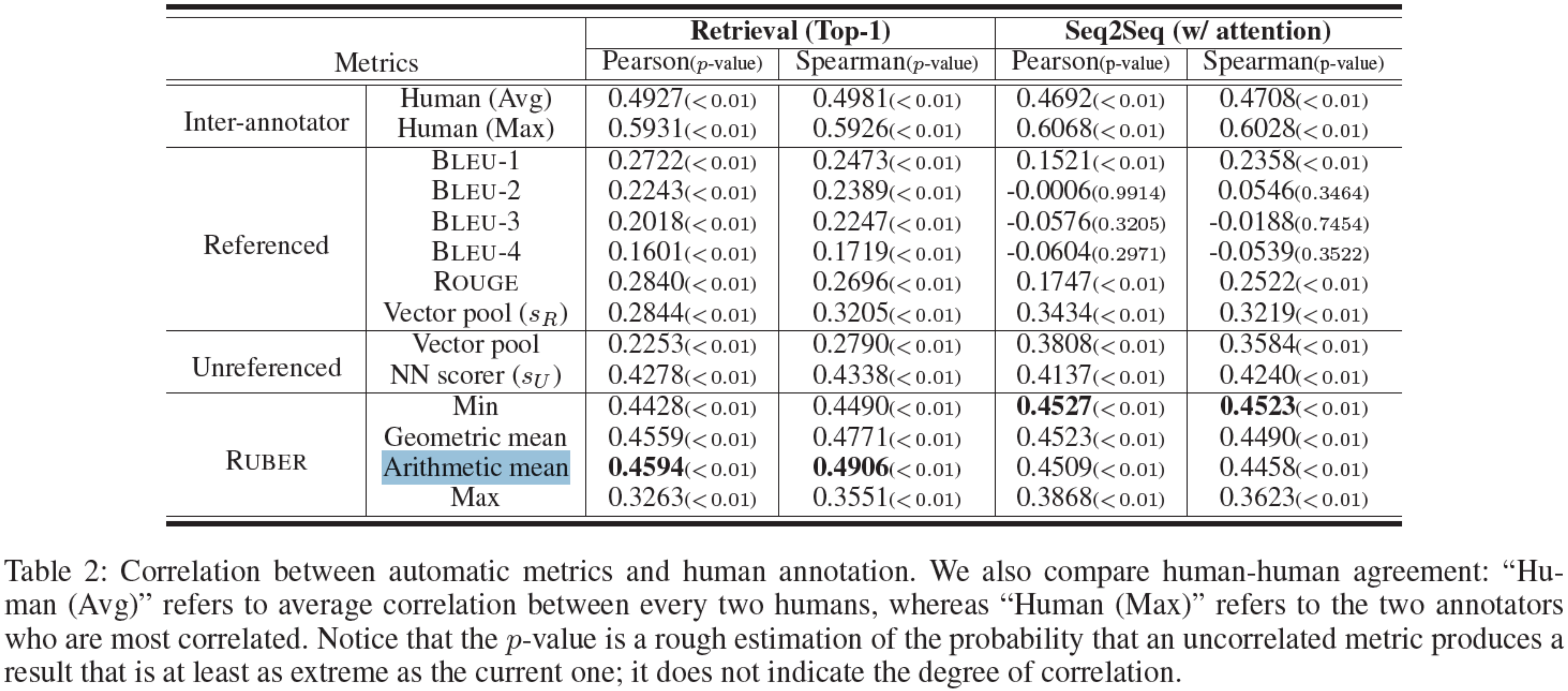
- Blending the normalized scores
  1. Min: $\min(\tilde{s}_R, \tilde{s}_U)$
  2. Max: $\mathrm{m}ax(\tilde{s}_R, \tilde{s}_U)$
  3. Geometric mean: $(\tilde{s}_R * \tilde{s}_U)^{1/2}$
  4. Arithmetic mean: $(\tilde{s}_R + \tilde{s}_U)/2$

$$\tilde{s}_R = \frac{s_R - \min(s_R)}{\max(s_R) - \min(s_R)}$$

$$\tilde{s}_U = \frac{s_U - \min(s_U)}{\max(s_U) - \min(s_U)}$$

RUBER-Experiments

- Dataset: Douban

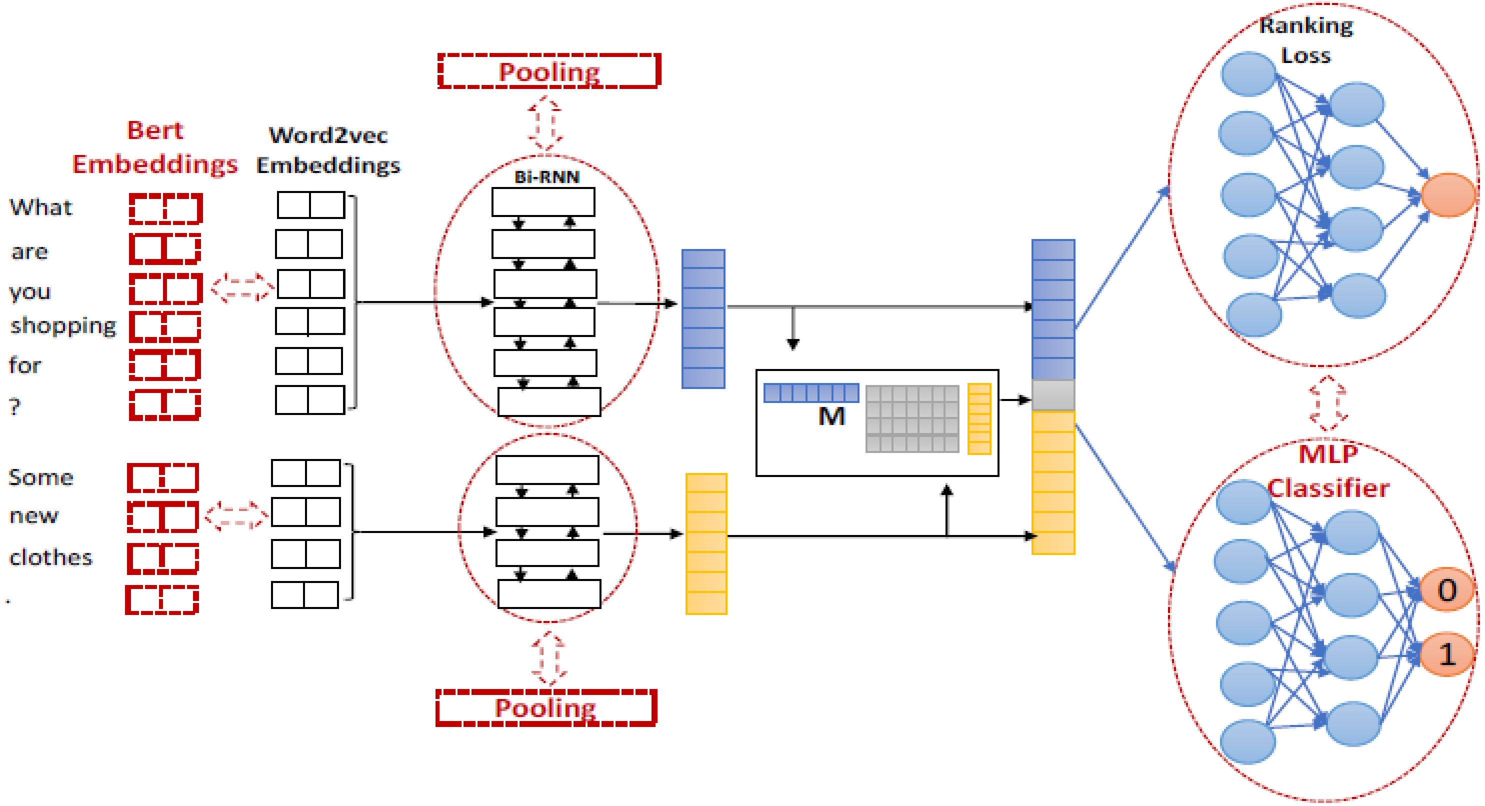


Metrics		Retrieval (Top-1)		Seq2Seq (w/ attention)	
		Pearson(p-value)	Spearman(p-value)	Pearson(p-value)	Spearman(p-value)
Inter-annotator	Human (Avg)	0.4927($<$0.01)	0.4981($<$0.01)	0.4692($<$0.01)	0.4708($<$0.01)
	Human (Max)	0.5931($<$0.01)	0.5926($<$0.01)	0.6068($<$0.01)	0.6028($<$0.01)
Referenced	BLEU-1	0.2722($<$0.01)	0.2473($<$0.01)	0.1521($<$0.01)	0.2358($<$0.01)
	BLEU-2	0.2243($<$0.01)	0.2389($<$0.01)	-0.0006(0.9914)	0.0546(0.3464)
	BLEU-3	0.2018($<$0.01)	0.2247($<$0.01)	-0.0576(0.3205)	-0.0188(0.7454)
	BLEU-4	0.1601($<$0.01)	0.1719($<$0.01)	-0.0604(0.2971)	-0.0539(0.3522)
	ROUGE	0.2840($<$0.01)	0.2696($<$0.01)	0.1747($<$0.01)	0.2522($<$0.01)
	Vector pool (s_R)	0.2844($<$0.01)	0.3205($<$0.01)	0.3434($<$0.01)	0.3219($<$0.01)
Unreferenced	Vector pool	0.2253($<$0.01)	0.2790($<$0.01)	0.3808($<$0.01)	0.3584($<$0.01)
	NN scorer (s_U)	0.4278($<$0.01)	0.4338($<$0.01)	0.4137($<$0.01)	0.4240($<$0.01)
RUBER	Min	0.4428($<$0.01)	0.4490($<$0.01)	**0.4527**($<$0.01)	**0.4523**($<$0.01)
	Geometric mean	0.4559($<$0.01)	0.4771($<$0.01)	0.4523($<$0.01)	0.4490($<$0.01)
	Arithmetic mean	**0.4594**($<$0.01)	**0.4906**($<$0.01)	0.4509($<$0.01)	0.4458($<$0.01)
	Max	0.3263($<$0.01)	0.3551($<$0.01)	0.3868($<$0.01)	0.3623($<$0.01)

Table 2: Correlation between automatic metrics and human annotation. We also compare human-human agreement: "Human (Avg)" refers to average correlation between every two humans, whereas "Human (Max)" refers to the two annotators who are most correlated. Notice that the p-value is a rough estimation of the probability that an uncorrelated metric produces a result that is at least as extreme as the current one; it does not indicate the degree of correlation.

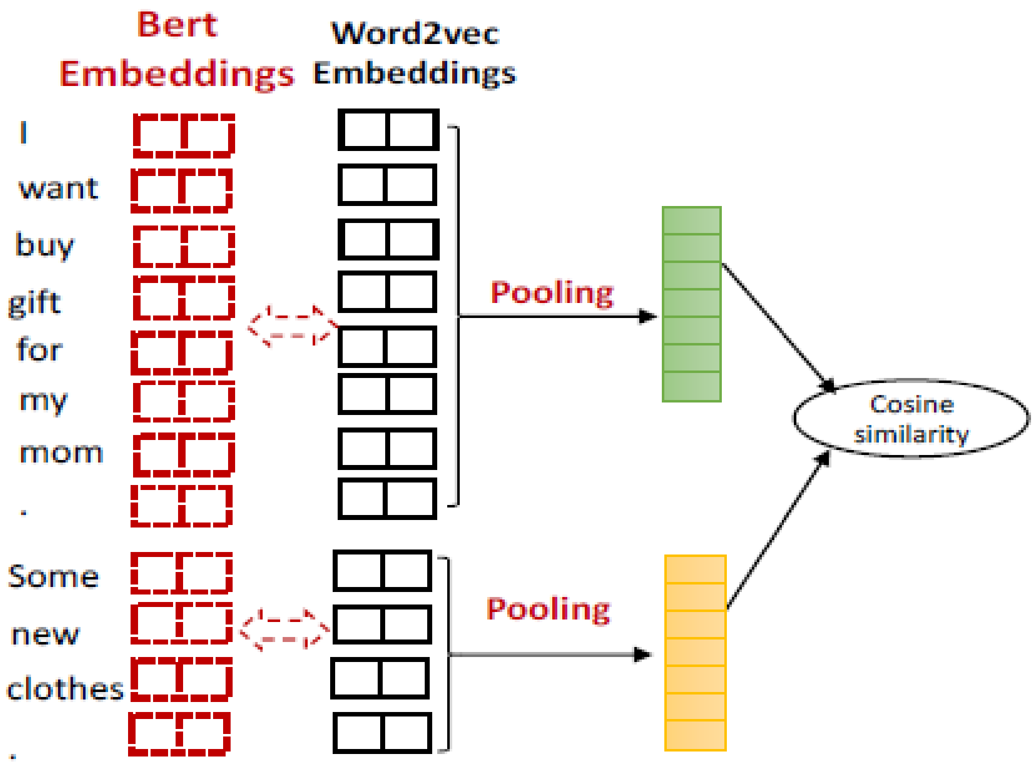
RUBER-An Extension with BERT

- Unreferenced Metric



RUBER-An Extension with BERT

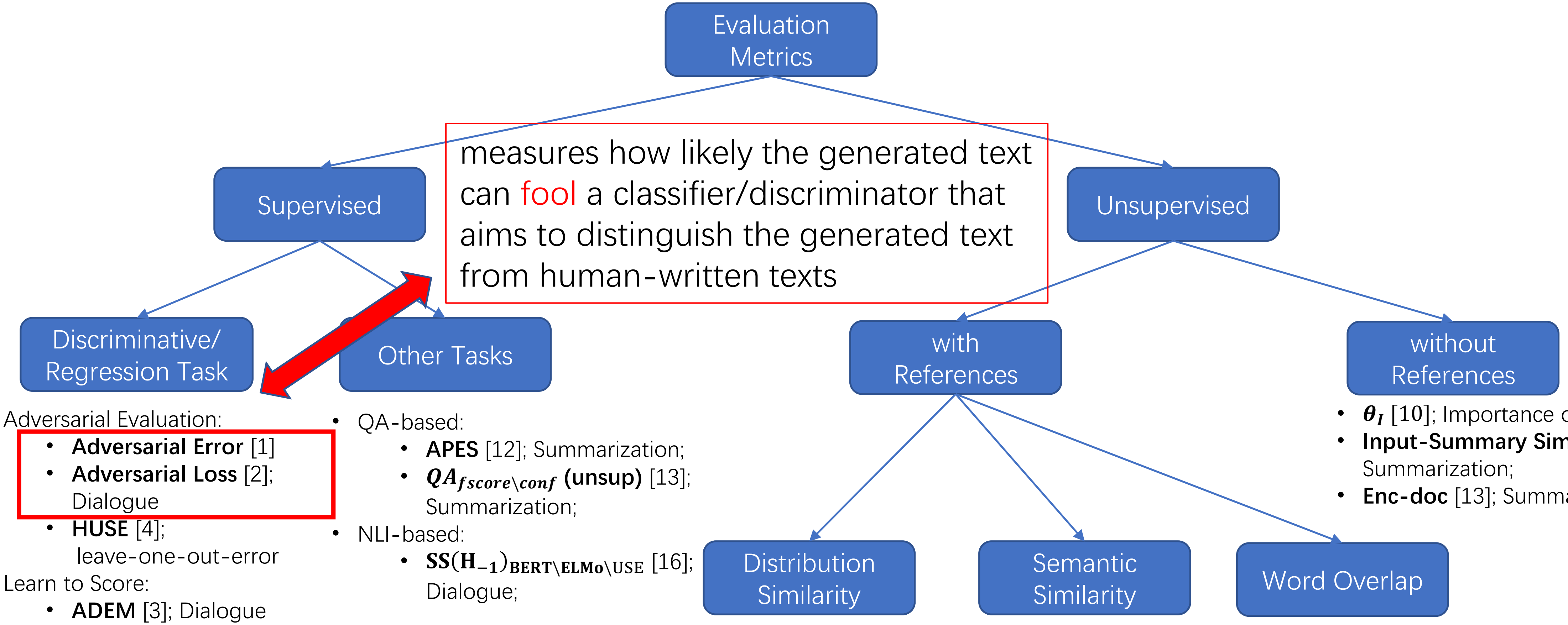
- Referenced Metric



RUBER-Conclusions

- A learnable, flexible hybrid metric for open-domain dialogue systems.
- Still supervised because of requiring training

Key Ideas of Other Metrics



Evaluation Metrics

Supervised

Unsupervised

Discriminative/ Regression Task

Other Tasks

with References

without References

Distribution Similarity

Semantic Similarity

Word Overlap

measures how likely the generated text can fool a classifier/discriminator that aims to distinguish the generated text from human-written texts

- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

- QA-based:
  - **APES** [12]; Summarization;
  - $QA_{fscore\backslash conf}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

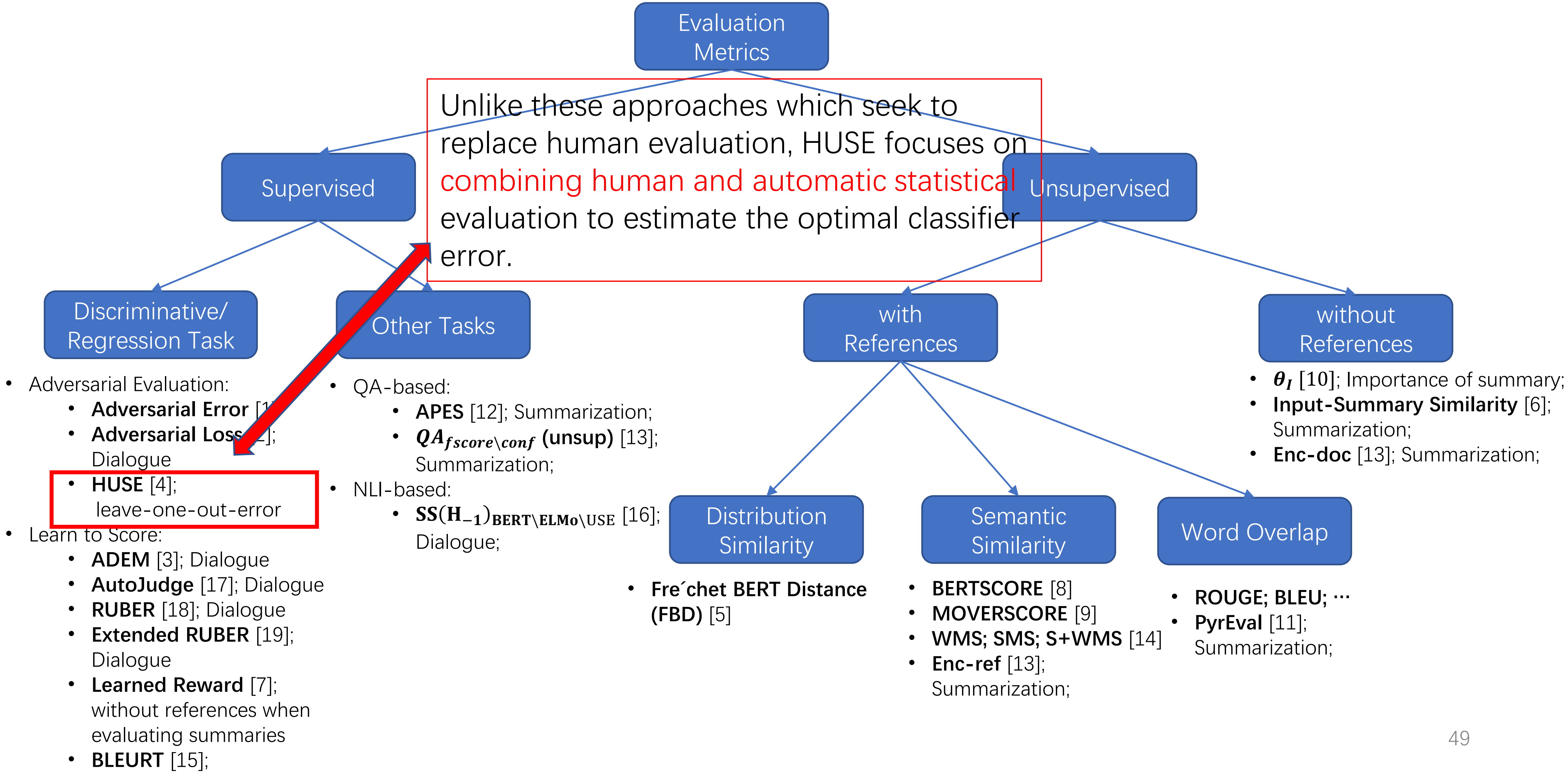
- **Fre´chet BERT Distance (FBD)** [5]

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

- **ROUGE; BLEU; …**
- **PyrEval** [11]; Summarization;

- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Key Ideas of Other Metrics



Unlike these approaches which seek to replace human evaluation, HUSE focuses on combining human and automatic statistical evaluation to estimate the optimal classifier error.

- Adversarial Evaluation:
  - **Adversarial Error** [1];
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

- QA-based:
  - **APES** [12]; Summarization;
  - $\boldsymbol{QA_{fscore\backslash conf}}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

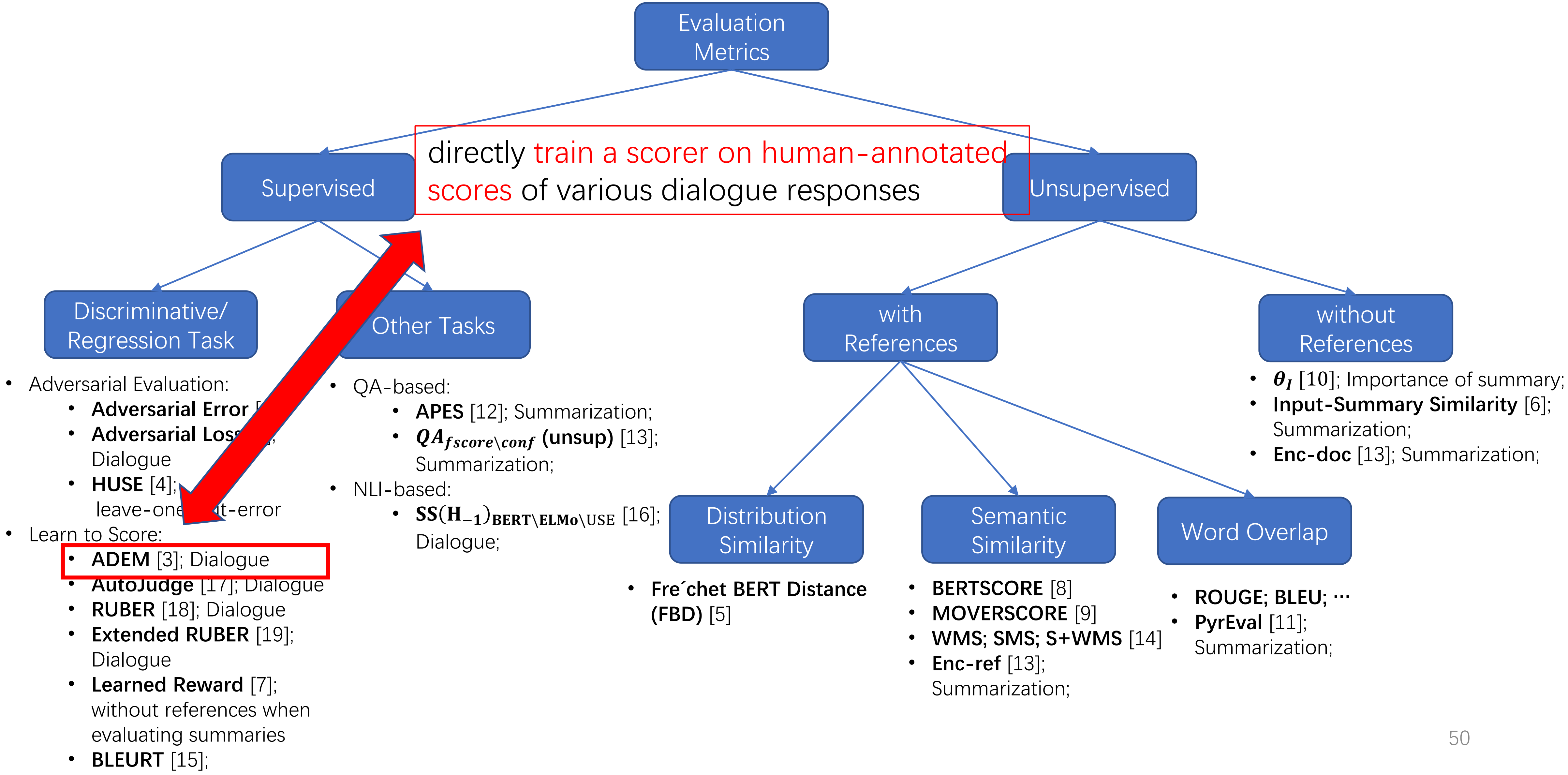
- **Fre´chet BERT Distance (FBD)** [5]

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

- **ROUGE; BLEU; ···**
- **PyrEval** [11]; Summarization;

- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Key Ideas of Other Metrics



Evaluation Metrics

Supervised

directly train a scorer on human-annotated scores of various dialogue responses

Unsupervised

Discriminative/ Regression Task

- Adversarial Evaluation:
  - **Adversarial Error** [illegible]
  - **Adversarial Loss** [illegible]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

Other Tasks

- QA-based:
  - **APES** [12]; Summarization;
  - $\boldsymbol{QA_{fscore\backslash conf}}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

with References

Distribution Similarity

- **Fre´chet BERT Distance (FBD)** [5]

Semantic Similarity

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

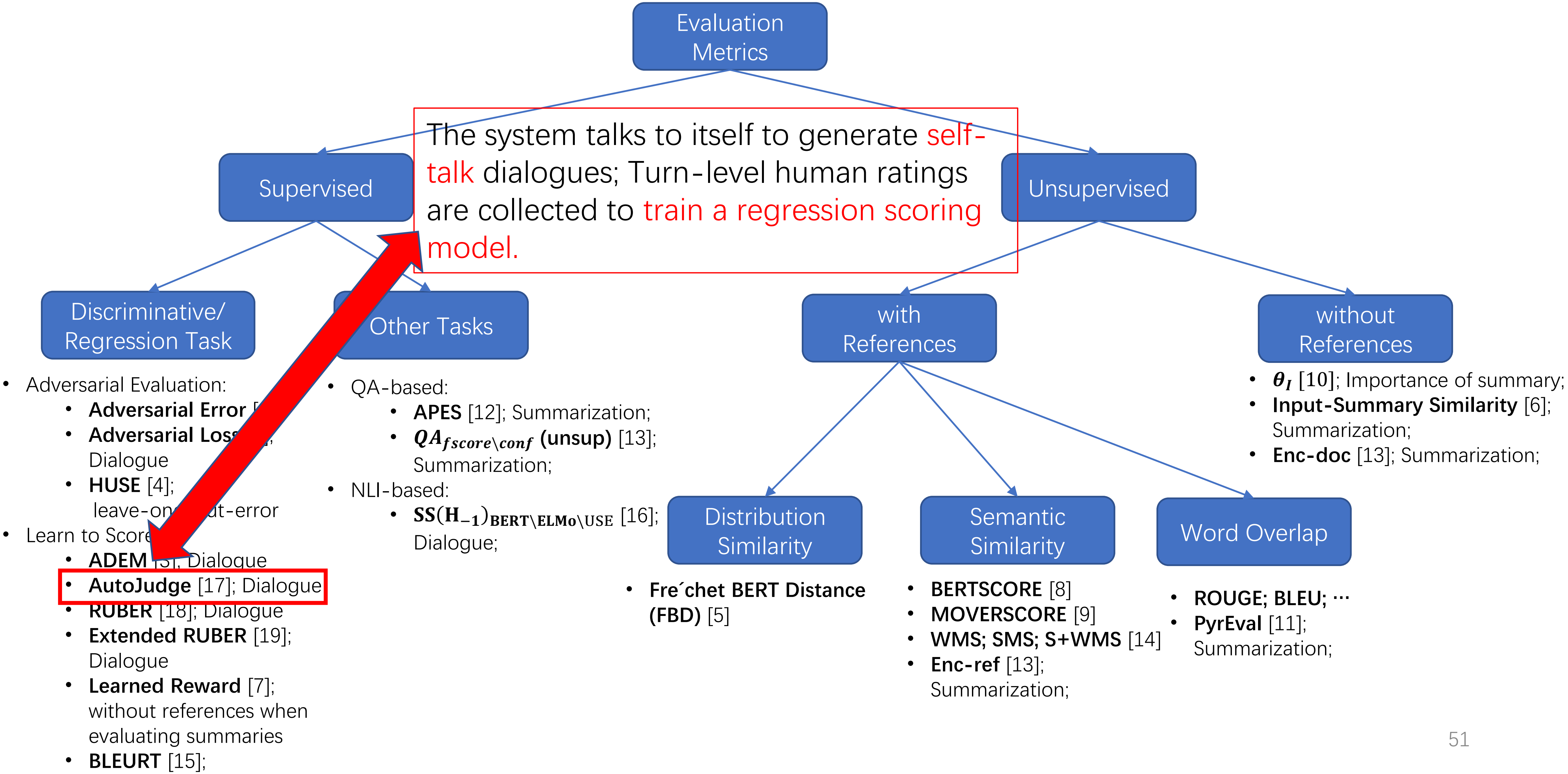
Word Overlap

- **ROUGE; BLEU; …**
- **PyrEval** [11]; Summarization;

without References

- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Key Ideas of Other Metrics



The system talks to itself to generate self-talk dialogues; Turn-level human ratings are collected to train a regression scoring model.

- Adversarial Evaluation:
  - **Adversarial Error** [
  - **Adversarial Loss** ; Dialogue
  - **HUSE** [4]; leave-on ut-error
- Learn to Score
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

- QA-based:
  - **APES** [12]; Summarization;
  - $QA_{fscore\backslash conf}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $SS(H_{-1})_{BERT\backslash ELMo\backslash USE}$ [16]; Dialogue;

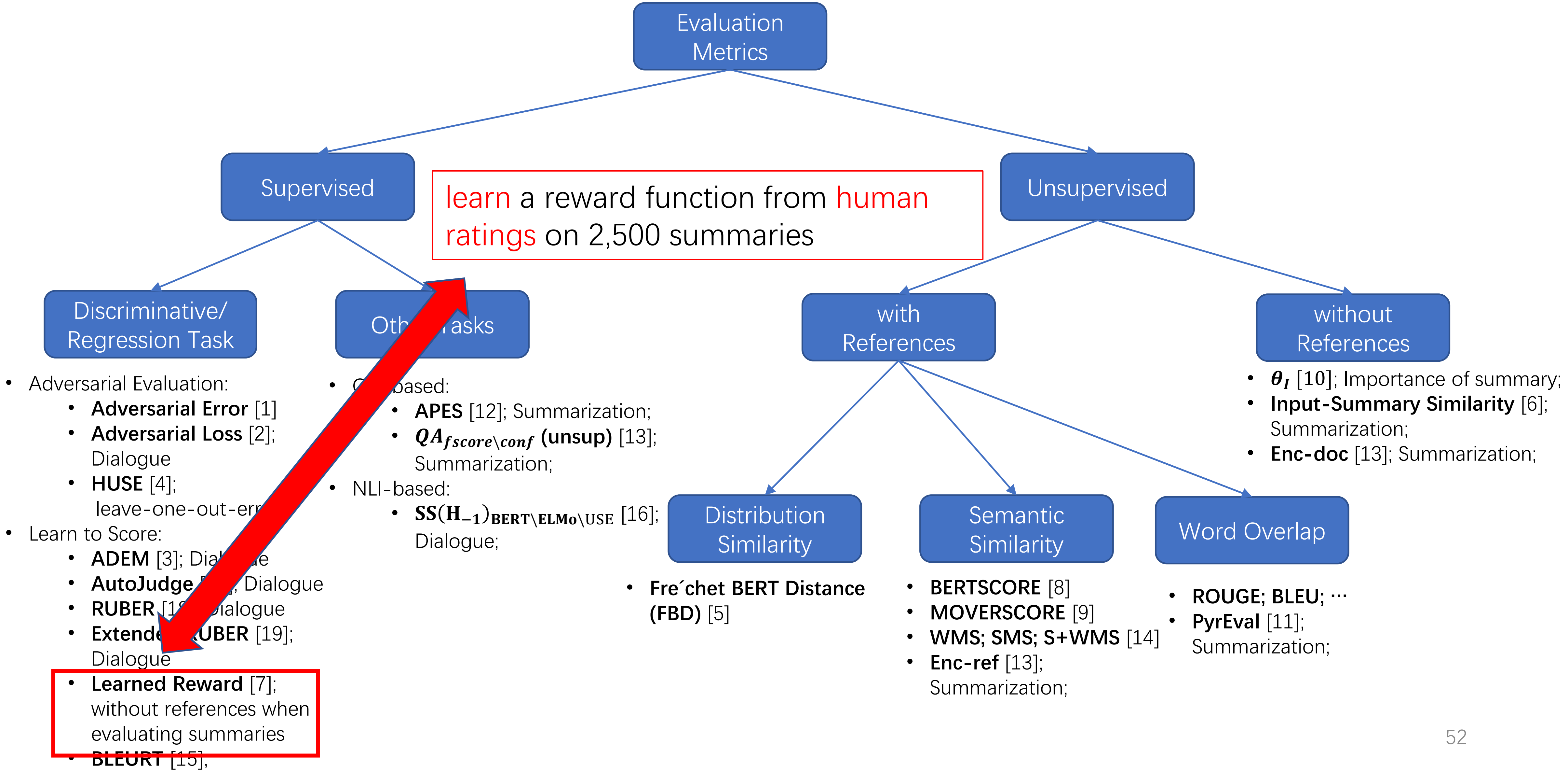
- **Fre´chet BERT Distance (FBD)** [5]

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

- **ROUGE; BLEU;** ···
- **PyrEval** [11]; Summarization;

- θ_I [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Key Ideas of Other Metrics



Evaluation Metrics

Supervised

Unsupervised

learn a reward function from human ratings on 2,500 summaries

Discriminative/ Regression Task

- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-err
- Learn to Score:
  - **ADEM** [3]; Dia e
  - **AutoJudge** , Dialogue
  - **RUBER** [1 ialogue
  - **Extende UBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

Oth asks

- Q based:
  - **APES** [12]; Summarization;
  - $\boldsymbol{QA_{fscore\backslash conf}}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

with References

Distribution Similarity

- **Fre´chet BERT Distance (FBD)** [5]

Semantic Similarity

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

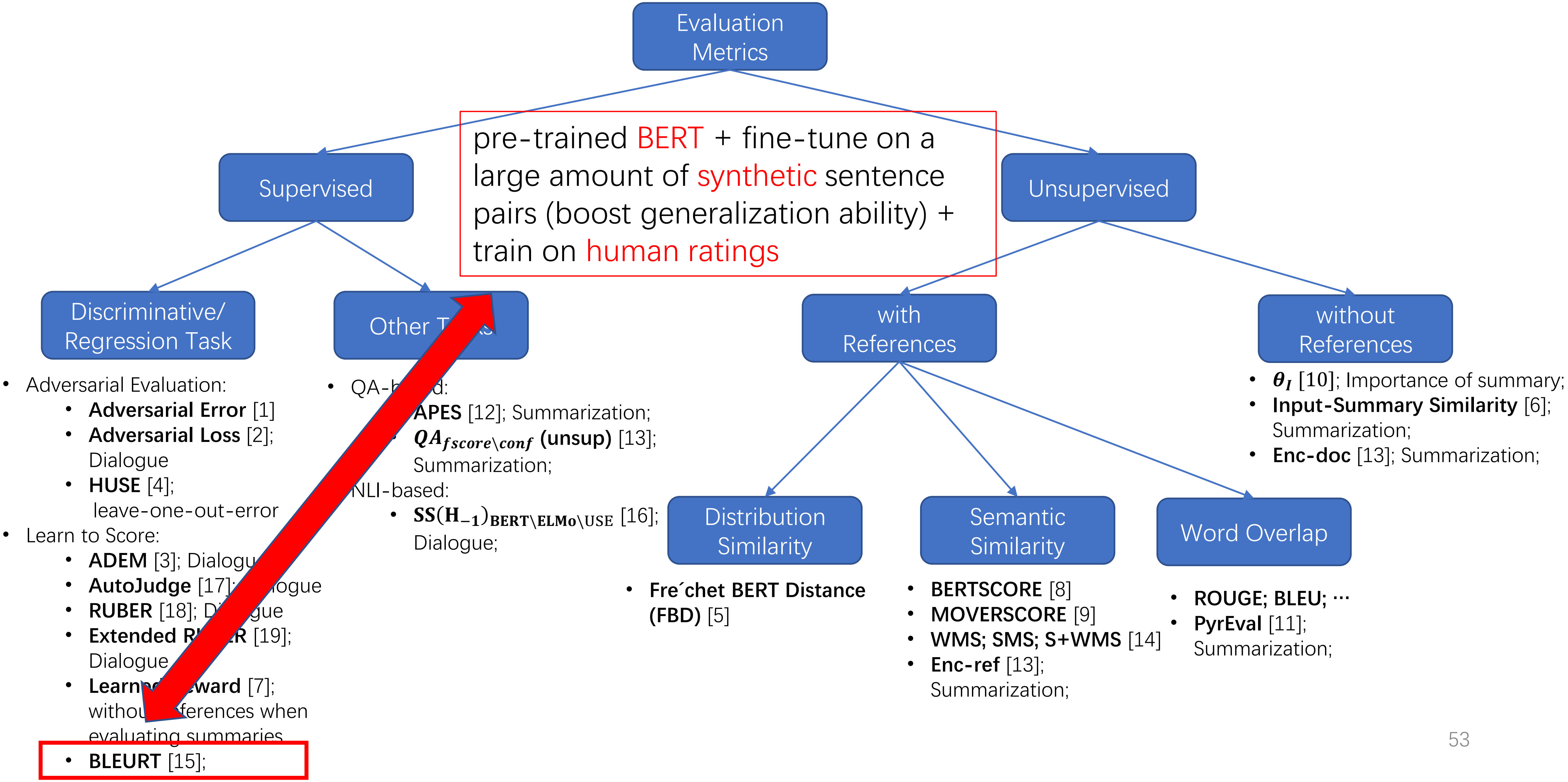
Word Overlap

- **ROUGE; BLEU; …**
- **PyrEval** [11]; Summarization;

without References

- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Key Ideas of Other Metrics



pre-trained BERT + fine-tune on a large amount of synthetic sentence pairs (boost generalization ability) + train on human ratings

- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

- QA-based:
  - **APES** [12]; Summarization;
  - $QA_{fscore\backslash conf}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

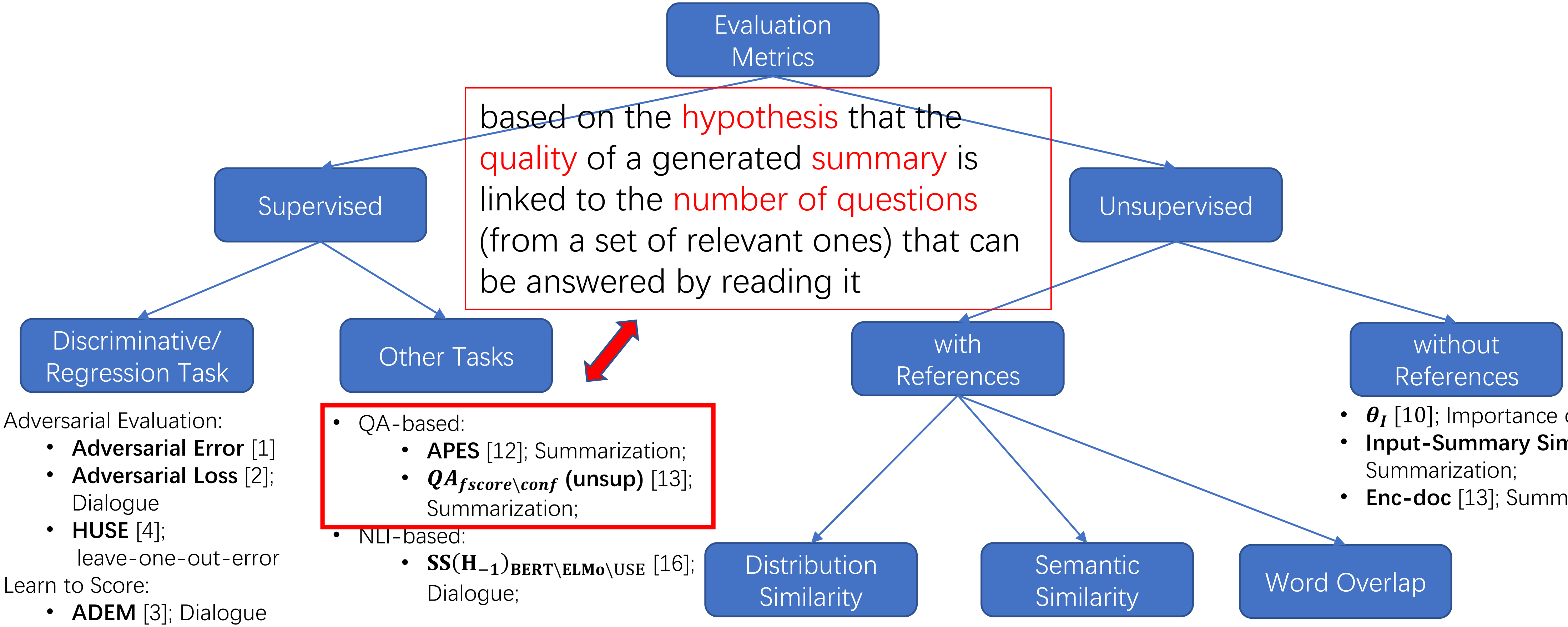
- **Fre´chet BERT Distance (FBD)** [5]

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

- **ROUGE; BLEU;** …
- **PyrEval** [11]; Summarization;

- θ_I [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Key Ideas of Other Metrics



based on the hypothesis that the quality of a generated summary is linked to the number of questions (from a set of relevant ones) that can be answered by reading it

- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

- QA-based:
  - **APES** [12]; Summarization;
  - $\boldsymbol{QA_{fscore\backslash conf}}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

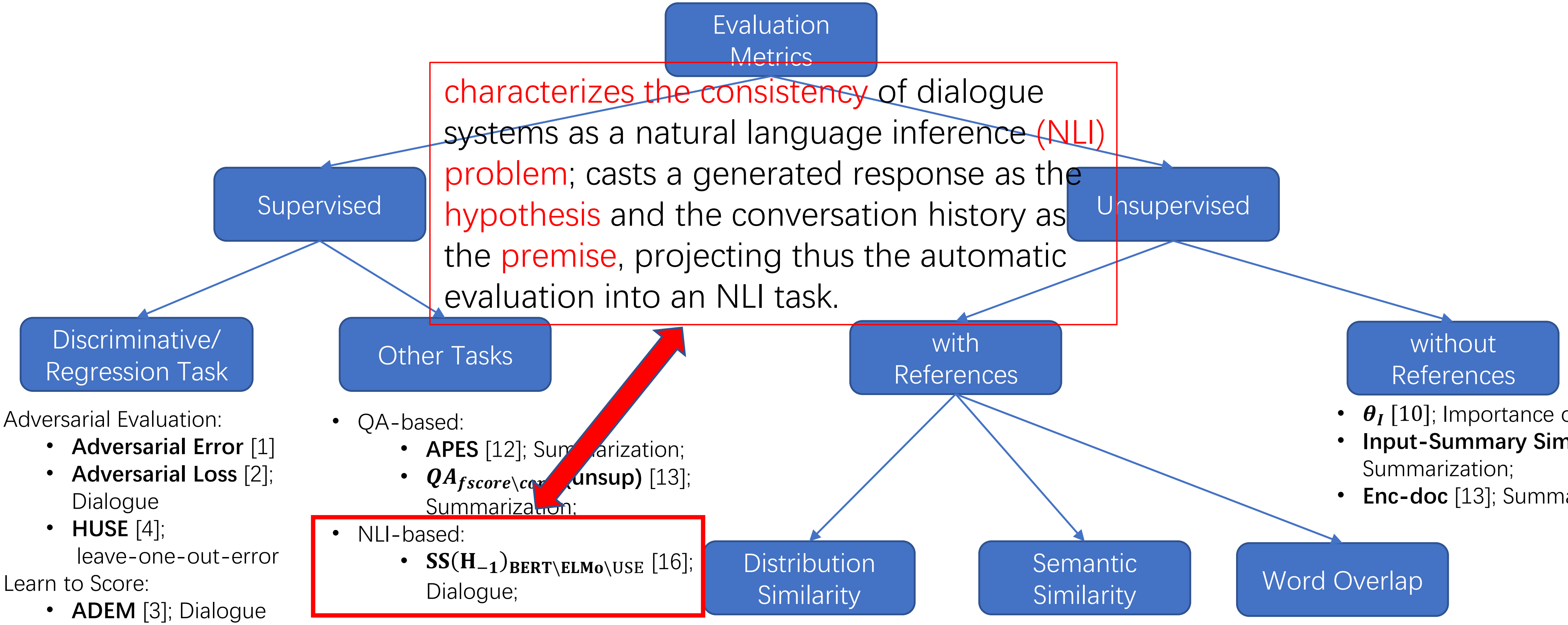
- **Fre´chet BERT Distance (FBD)** [5]

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

- **ROUGE; BLEU; ⋯**
- **PyrEval** [11]; Summarization;

- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Key Ideas of Other Metrics



characterizes the consistency of dialogue systems as a natural language inference (NLI) problem; casts a generated response as the hypothesis and the conversation history as the premise, projecting thus the automatic evaluation into an NLI task.

- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

- QA-based:
  - **APES** [12]; Summarization;
  - $\boldsymbol{QA_{fscore\backslash co}}$**unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

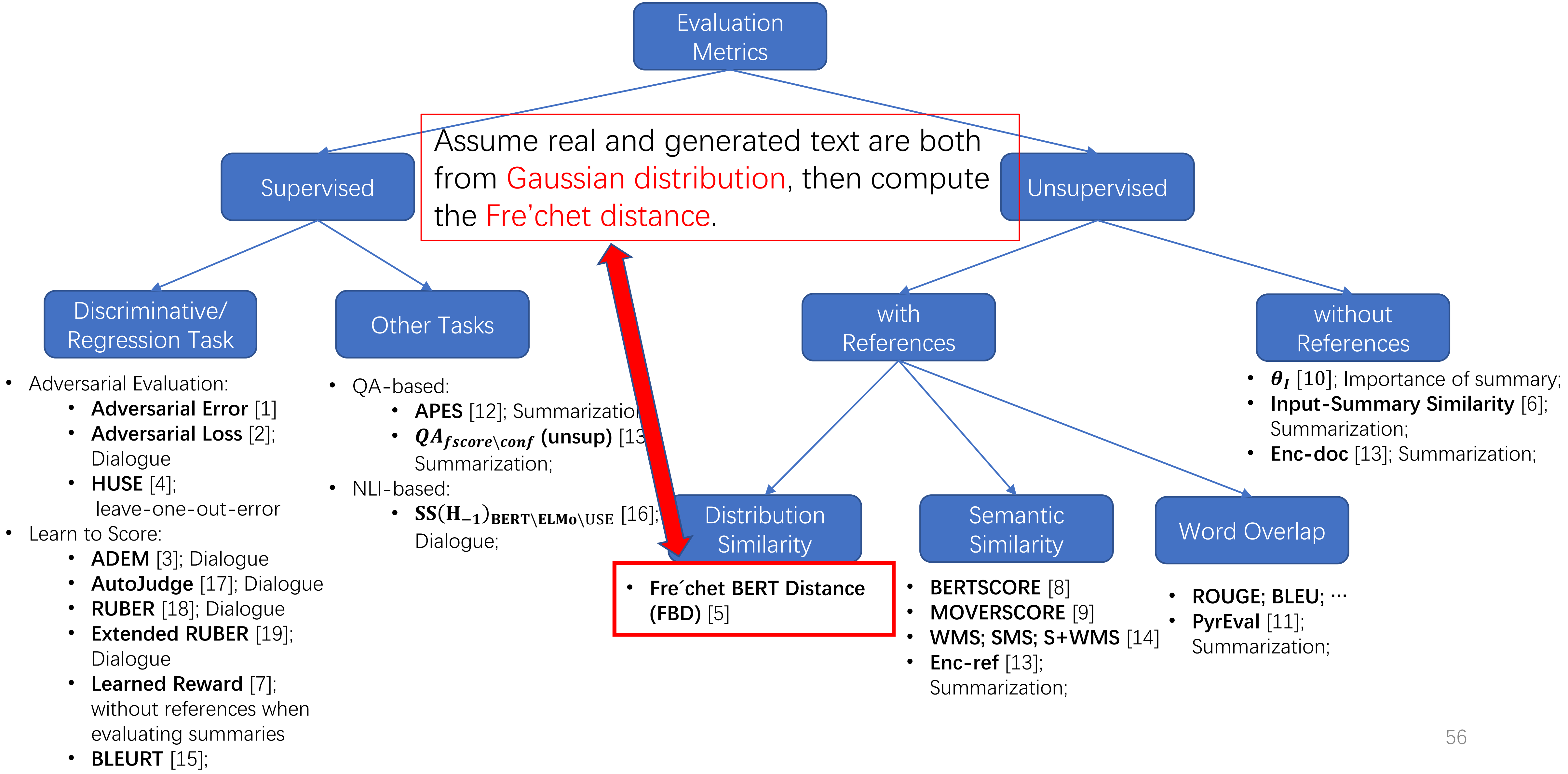
- **Fre´chet BERT Distance (FBD)** [5]

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

- **ROUGE; BLEU; …**
- **PyrEval** [11]; Summarization;

- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Key Ideas of Other Metrics



Evaluation Metrics

Assume real and generated text are both from Gaussian distribution, then compute the Fre’chet distance.

Supervised

Discriminative/ Regression Task

- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

Other Tasks

- QA-based:
  - **APES** [12]; Summarizatio
  - $\boldsymbol{QA_{fscore\backslash conf}}$ **(unsup)** [13 Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

Unsupervised

with References

Distribution Similarity

- **Fre´chet BERT Distance (FBD)** [5]

Semantic Similarity

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

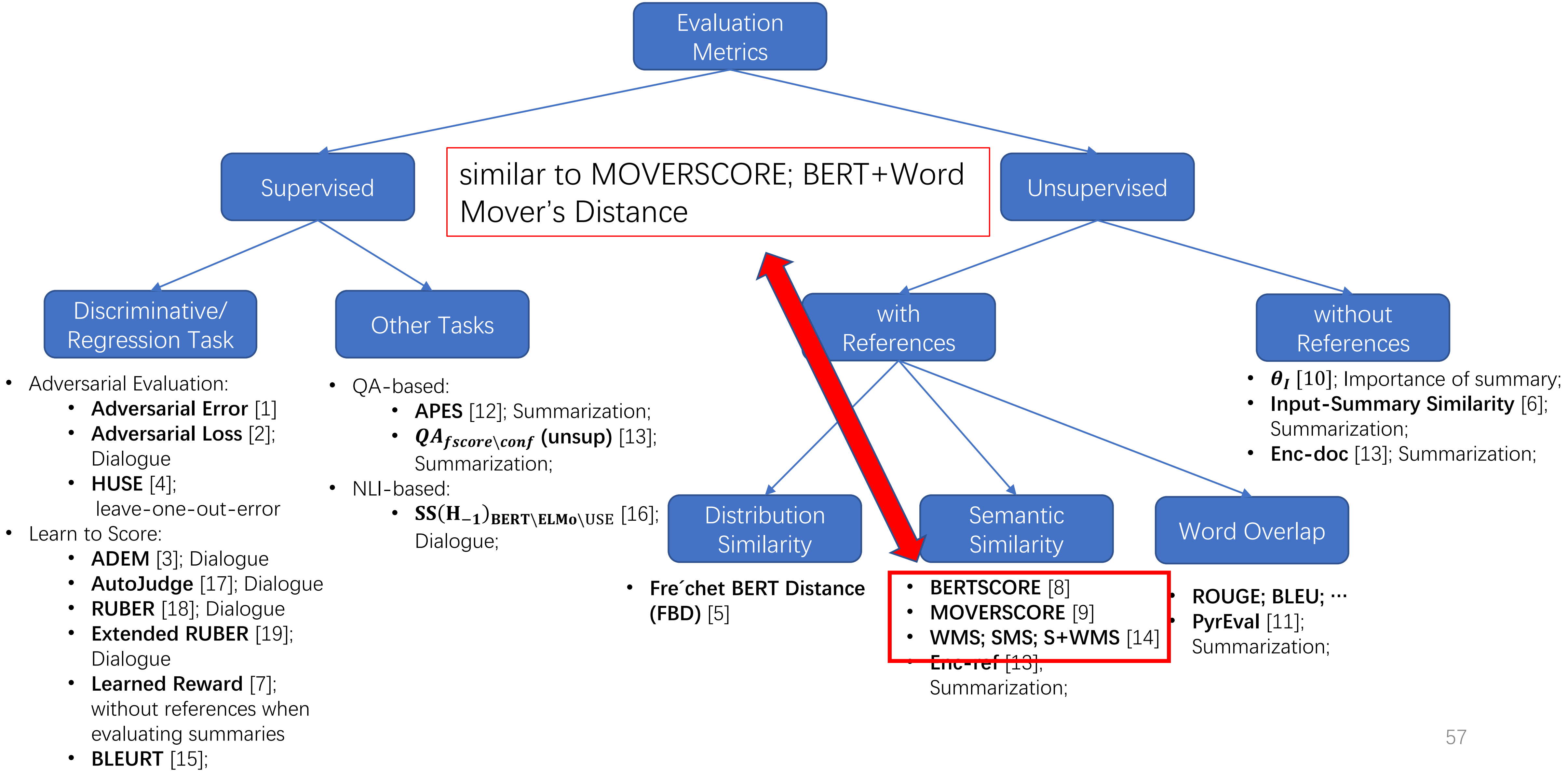
Word Overlap

- **ROUGE; BLEU;** ···
- **PyrEval** [11]; Summarization;

without References

- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Key Ideas of Other Metrics



Evaluation Metrics

Supervised

similar to MOVERSCORE; BERT+Word Mover's Distance

Unsupervised

Discriminative/ Regression Task

- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

Other Tasks

- QA-based:
  - **APES** [12]; Summarization;
  - $QA_{fscore\backslash conf}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

with References

Distribution Similarity

- **Fre´chet BERT Distance (FBD)** [5]

Semantic Similarity

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

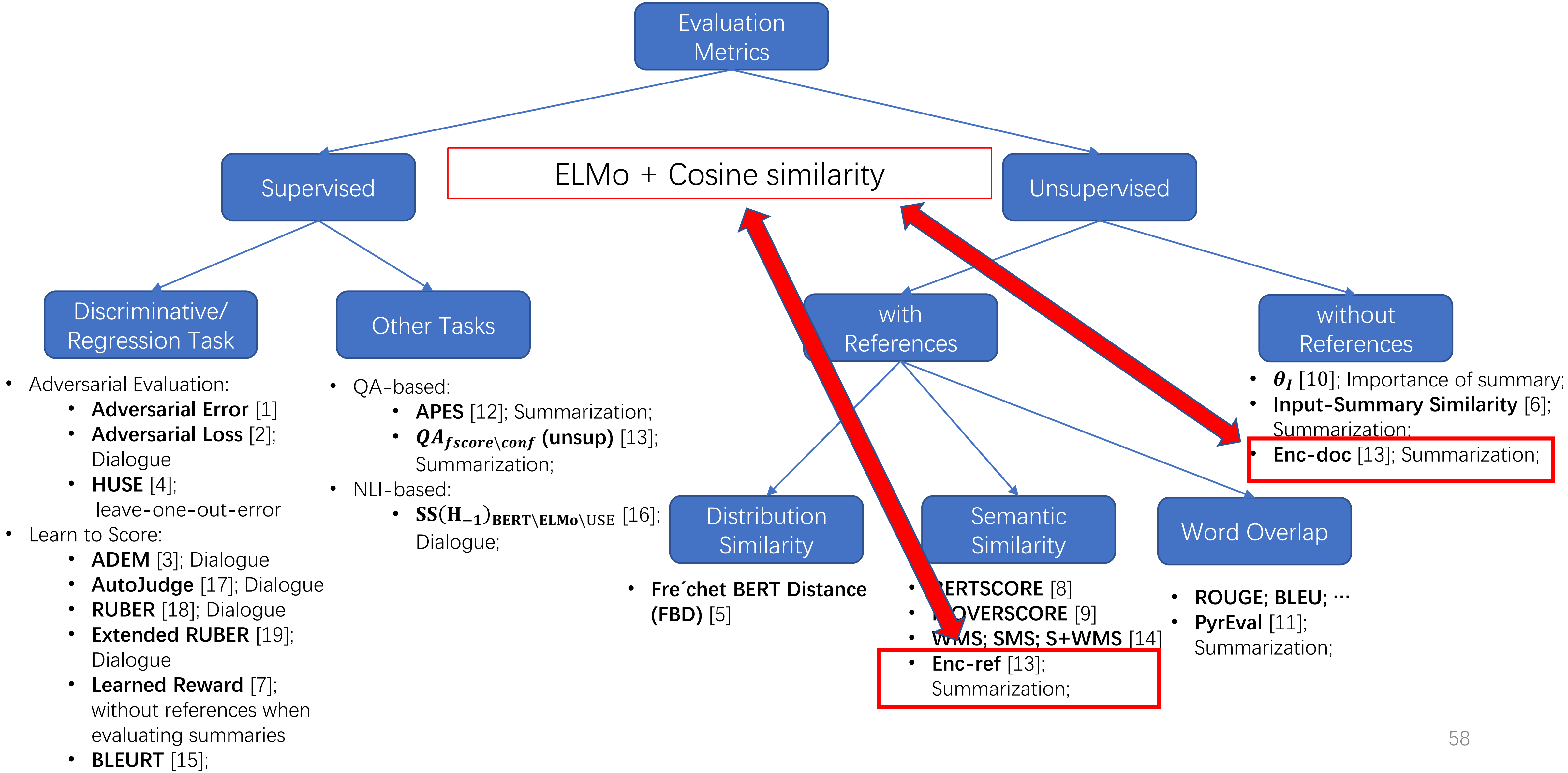
Word Overlap

- **ROUGE; BLEU; …**
- **PyrEval** [11]; Summarization;

without References

- θ_I [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Key Ideas of Other Metrics



Evaluation Metrics

Supervised

ELMo + Cosine similarity

Unsupervised

Discriminative/ Regression Task

- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

Other Tasks

- QA-based:
  - **APES** [12]; Summarization;
  - $\boldsymbol{QA_{fscore\backslash conf}}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

with References

Distribution Similarity

- **Fre´chet BERT Distance (FBD)** [5]

Semantic Similarity

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

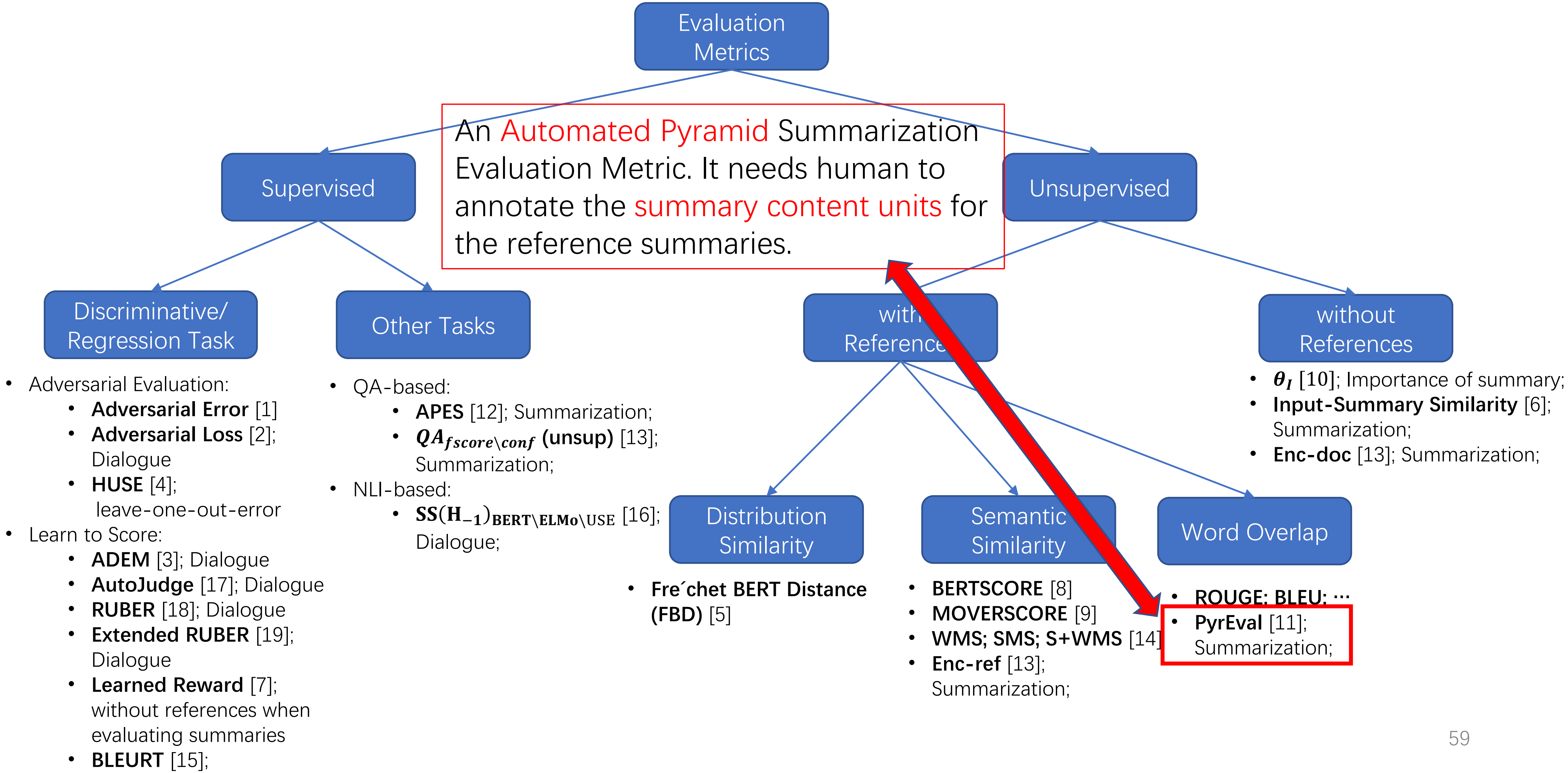
Word Overlap

- **ROUGE; BLEU; …**
- **PyrEval** [11]; Summarization;

without References

- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Key Ideas of Other Metrics



Evaluation Metrics

An <span>Automated Pyramid</span> Summarization Evaluation Metric. It needs human to annotate the summary content units for the reference summaries.

Supervised

Unsupervised

Discriminative/ Regression Task

- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

Other Tasks

- QA-based:
  - **APES** [12]; Summarization;
  - $\boldsymbol{QA_{fscore\backslash conf}}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

with References

Distribution Similarity

- **Fre´chet BERT Distance (FBD)** [5]

Semantic Similarity

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

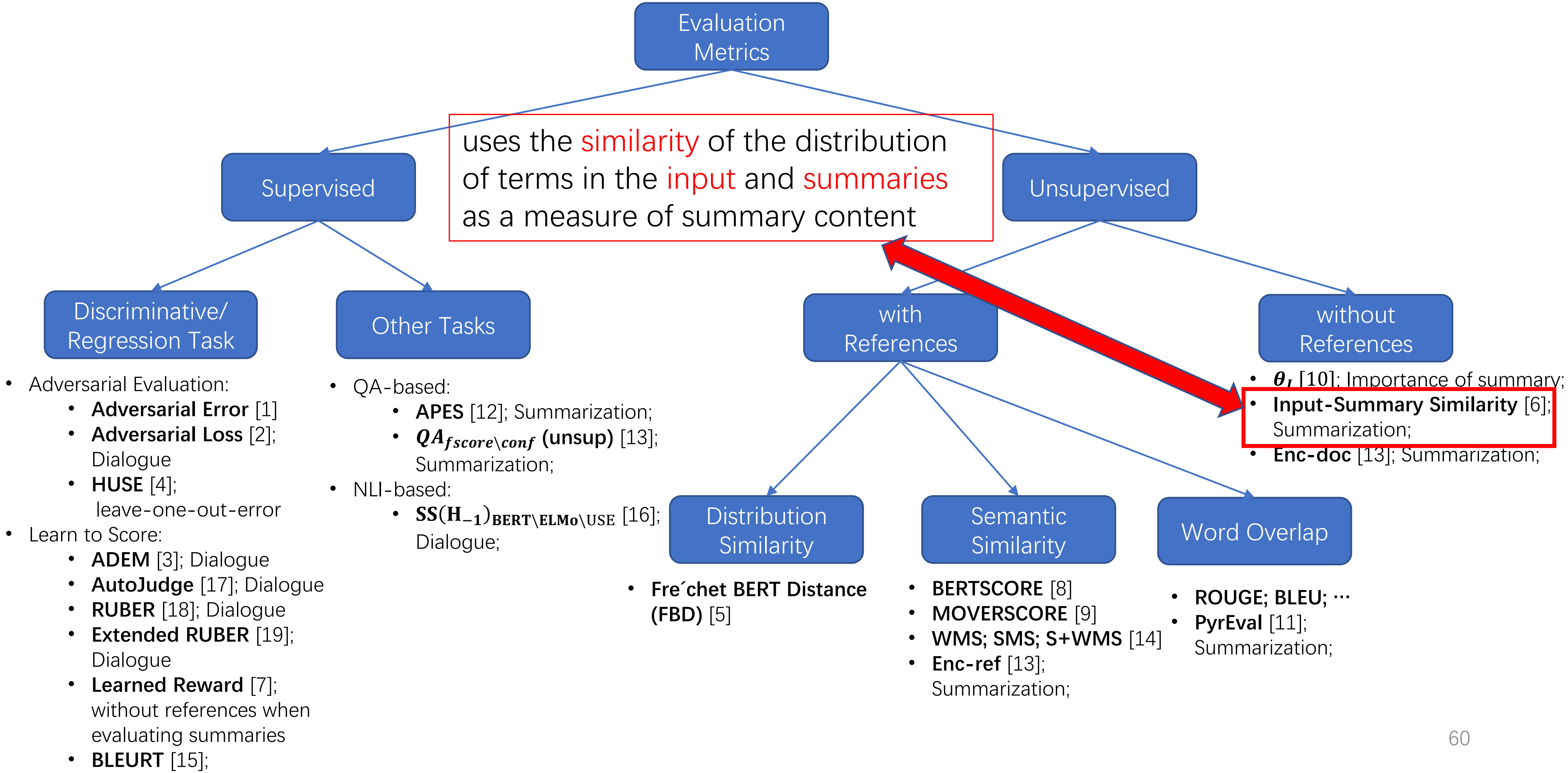
Word Overlap

- **ROUGE; BLEU; …**
- **PyrEval** [11]; Summarization;

without References

- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Key Ideas of Other Metrics



uses the similarity of the distribution of terms in the input and summaries as a measure of summary content

Evaluation Metrics

Supervised

Unsupervised

Discriminative/ Regression Task

- Adversarial Evaluation:
  - **Adversarial Error** [1]
  - **Adversarial Loss** [2]; Dialogue
  - **HUSE** [4]; leave-one-out-error
- Learn to Score:
  - **ADEM** [3]; Dialogue
  - **AutoJudge** [17]; Dialogue
  - **RUBER** [18]; Dialogue
  - **Extended RUBER** [19]; Dialogue
  - **Learned Reward** [7]; without references when evaluating summaries
  - **BLEURT** [15];

Other Tasks

- QA-based:
  - **APES** [12]; Summarization;
  - $\boldsymbol{QA_{fscore\backslash conf}}$ **(unsup)** [13]; Summarization;
- NLI-based:
  - $\mathbf{SS(H_{-1})_{BERT\backslash ELMo\backslash USE}}$ [16]; Dialogue;

with References

Distribution Similarity

- **Fre´chet BERT Distance (FBD)** [5]

Semantic Similarity

- **BERTSCORE** [8]
- **MOVERSCORE** [9]
- **WMS; SMS; S+WMS** [14]
- **Enc-ref** [13]; Summarization;

Word Overlap

- **ROUGE; BLEU; ⋯**
- **PyrEval** [11]; Summarization;

without References

- $\boldsymbol{\theta_I}$ [10]; Importance of summary;
- **Input-Summary Similarity** [6]; Summarization;
- **Enc-doc** [13]; Summarization;

Conclusions

- We introduced a new metric for general text generation, summarization, and dialogue generation respectively.

- We briefly introduced the key ideas of various metrics based on the taxonomy.

- Unsupervised, semantic similarity based metrics are worthwhile to be engaged in your work.

References

- [1]. Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 10–21.
- [2]. Anjuli Kannan and Oriol Vinyals. 2017. Adversarial evaluation of dialogue models. arXiv preprint arXiv:1701.08198.
- [3]. Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1116–1126.
- [4]. Hashimoto, Tatsunori B., Hugh Zhang, and Percy Liang. "Unifying human and statistical evaluation for natural language generation." arXiv preprint arXiv:1904.02792 (2019).
- [5]. Montahaei, Ehsan, Danial Alihosseini, and Mahdieh Soleymani Baghshah. "Jointly measuring diversity and quality in text generation models." arXiv preprint arXiv:1904.03971 (2019).
- [6]. Louis, Annie, and Ani Nenkova. "Automatically assessing machine summary content without a gold standard." Computational Linguistics 39.2 (2013): 267-300.
- [7]. Böhm, Florian, et al. "Better rewards yield better summaries: Learning to summarise without references." arXiv preprint arXiv:1909.01214 (2019).
- [8]. Zhang, Tianyi, et al. "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675 (2019).
- [9]. Zhao, Wei, et al. "Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance." arXiv preprint arXiv:1909.02622 (2019).
- [10]. Peyrard, Maxime. "A simple theoretical model of importance for summarization." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

References

- [11]. Gao, Yanjun, Chen Sun, and Rebecca J. Passonneau. "Automated Pyramid Summarization Evaluation." Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). 2019.
- [12]. Eyal, Matan, Tal Baumel, and Michael Elhadad. "Question answering as an automatic evaluation metric for news article summarization." arXiv preprint arXiv:1906.00318 (2019).
- [13]. Sun, Simeng, and Ani Nenkova. "The Feasibility of Embedding Based Automatic Evaluation for Single Document Summarization." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.
- [14]. Clark, Elizabeth, Asli Celikyilmaz, and Noah A. Smith. "Sentence mover's similarity: Automatic evaluation for multi-sentence texts." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
- [15]. Sellam, Thibault, Dipanjan Das, and Ankur P. Parikh. "BLEURT: Learning Robust Metrics for Text Generation." arXiv preprint arXiv:2004.04696 (2020).
- [16]. Dziri, Nouha, et al. "Evaluating coherence in dialogue systems using entailment." arXiv preprint arXiv:1904.03371 (2019).
- [17]. Deriu, Jan, and Mark Cieliebak. "Towards a metric for automated conversational dialogue system evaluation and improvement." arXiv preprint arXiv:1909.12066 (2019).
- [18]. Tao, Chongyang, et al. "Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
- [19]. Ghazarian, Sarik, et al. "Better automatic evaluation of open-domain dialogue systems with contextualized embeddings." arXiv preprint arXiv:1904.10635 (2019).