自然語言處理與應用
Natural Language Processing and Applications

NLG Evaluations (Learning-based)

Instructor: 林英嘉 (Ying-Jia Lin)
2025/05/26

Course GitHub

Slido # NLP_0526

Phases for Building a Machine Learning Model

Data-preprocessing

↓

Model Training

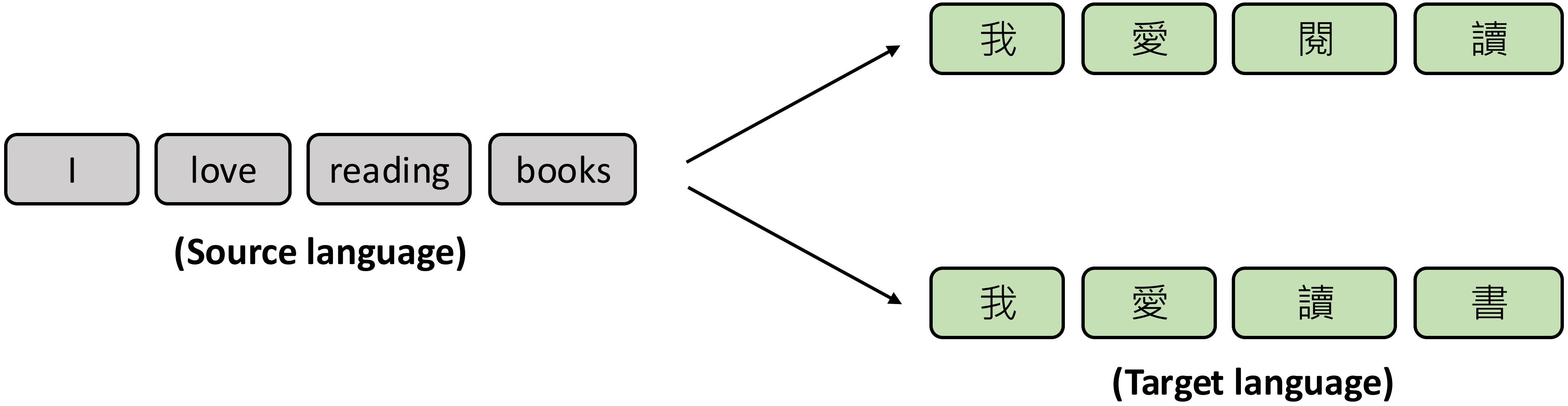
↓

Evaluations --→ Metrics --→

[Recap] How to evaluate natural language generation?

- Natural language is hard to evaluate due to <u>subjectivity</u> and language <u>diversity</u>.

For example: Machine Translation



- Human evaluations
- Automatic evaluations (We will focus on this topic.)

Medical Report Example

Ground Truth Report: the lungs are hyperexpanded consistent with emphysema the heart size and pulmonary vascularity appear within normal limits no pneumothorax or pleural effusion is seen patchy airspace disease is present in the right middle lobe degenerative changes are present spine

There is hyperexpansion of the lungs indicating emphysema. The heart and pulmonary vessels are within normal limits. No signs of pneumothorax or pleural effusion. Airspace disease affects the right middle lobe. Degenerative spinal changes are noted.

Evaluations

- Perplexity
- BLEU Score
- ROUGE Score

Rule-based (model-free)

- BERTScore
- BLEURT
- Mauve

Learning-based (model-based)

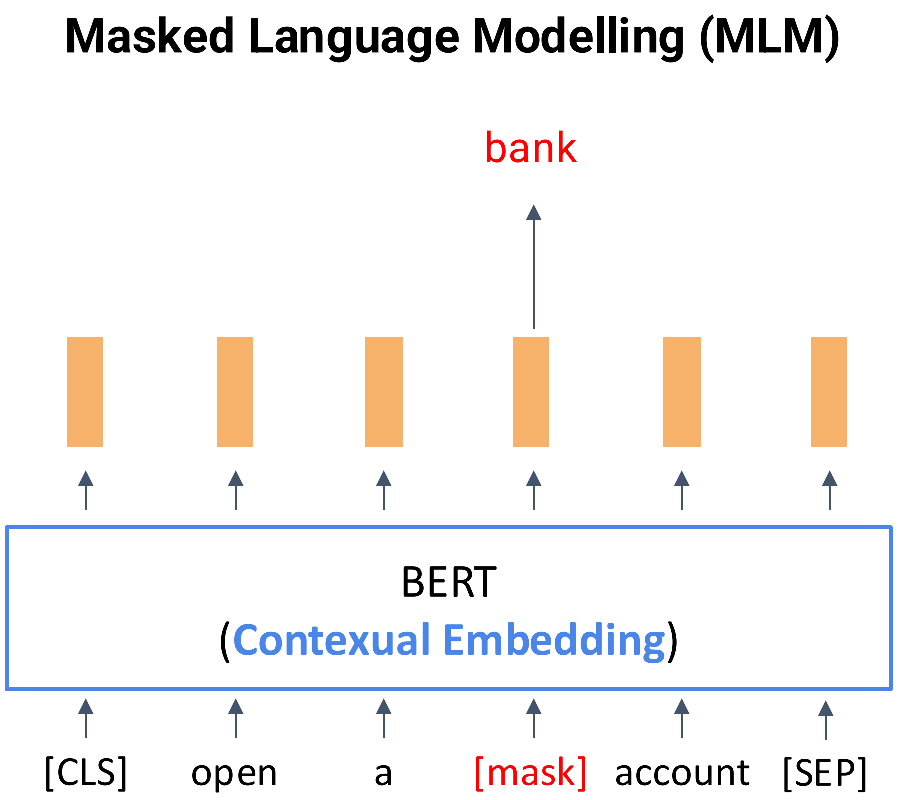
Issue of BLEU and ROUGE

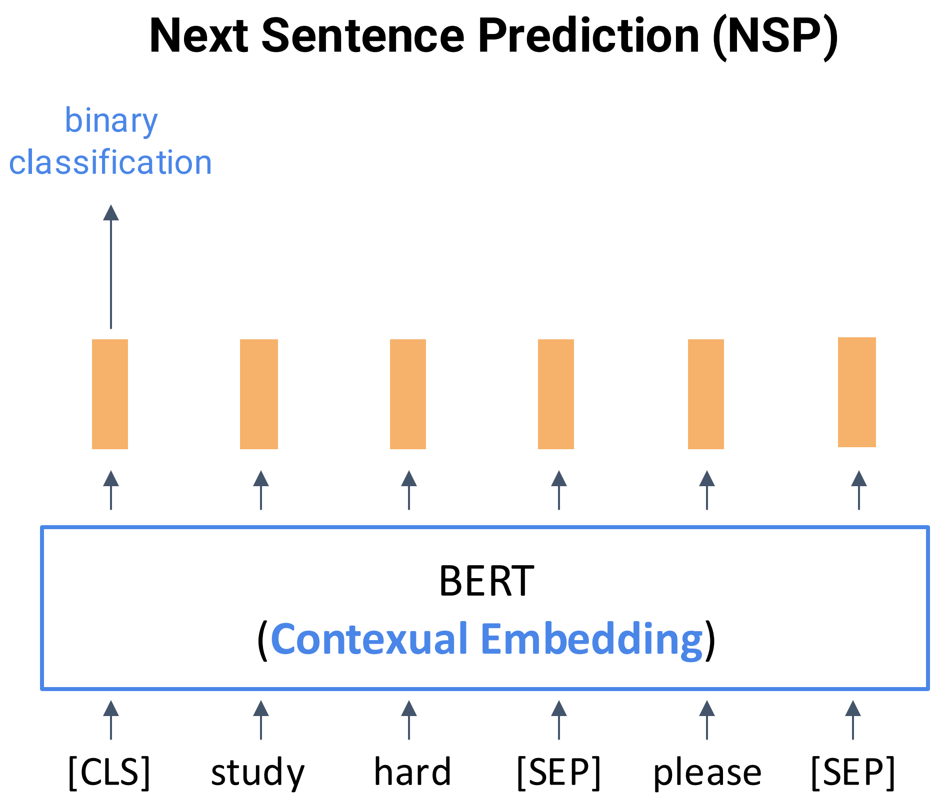
- Cons: Cannot meet language diversity
  - This mainly comes from the way for measuring overlapping rates.
- **Question**: Can we create an automatic metric to fix the issue?

- Next, we are going to introduce two learned automatic evaluation metrics
  - BERTScore (ICLR 2020)
  - BLEURT (ACL 2020)

(Recap) BERT: Bidirectional Encoder Representations from Transformers

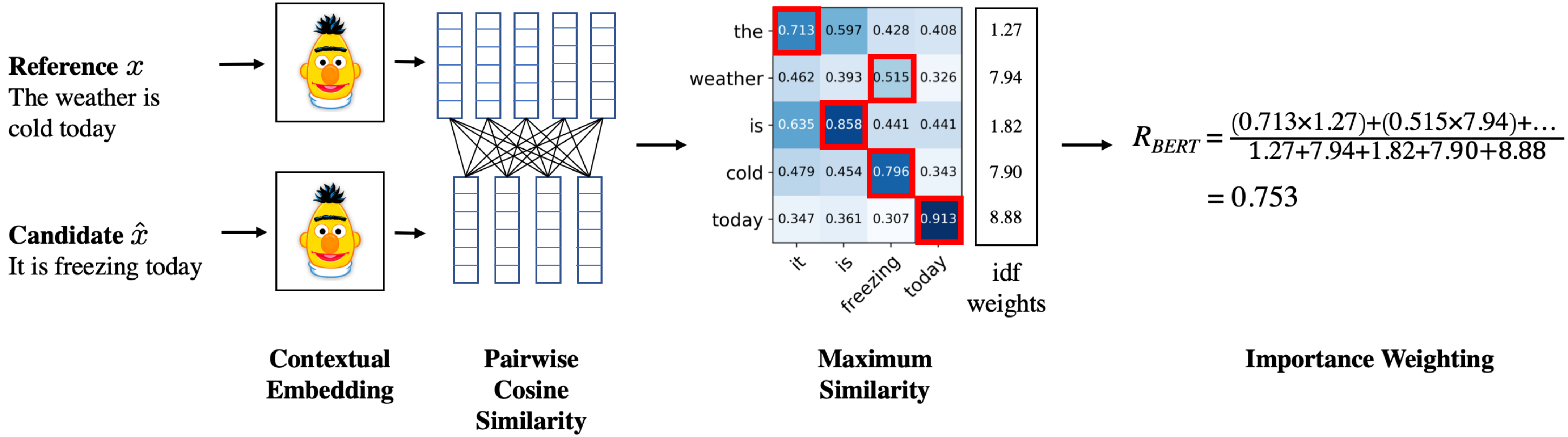
BERT was pre-trained with MLM and NSP objectives.





Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL 2019.

BERTScore – Overview



Zhang, Tianyi, et al. "BERTScore: Evaluating Text Generation with BERT." International Conference on Learning Representations. 2020.

BERTScore 使用範例

How's the weather today?

It is freezing today.

Candidate

The weather is cold today.

Reference

BERTScore → **95%**

代表生成的結果與
標準答案的語意相似性

BERTScore – Steps

Step 0: Prepare Reference x, Candidate $\hat{x}$, and a pre-trained BERT model

Step 1: Infer x and $\hat{x}$ with BERT respectively, get a sequence of output vectors $\langle \mathrm{x}_1, \ldots, \mathrm{x}_k \rangle$ for x and a sequence of output vectors $\langle \hat{\mathrm{x}}_1, \ldots, \hat{\mathrm{x}}_k \rangle$ for $\hat{x}$

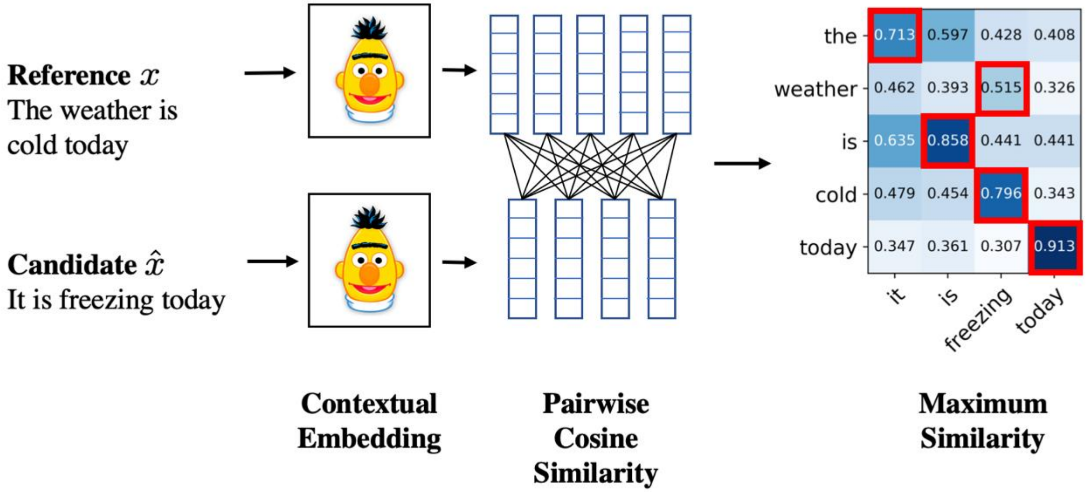
Zhang, Tianyi, et al. "BERTScore: Evaluating Text Generation with BERT." International Conference on Learning Representations. 2020.

BERTScore – Steps

Zhang, Tianyi, et al. "BERTScore: Evaluating Text Generation with BERT." International Conference on Learning Representations. 2020.

Step 2: Measure pairwise cosine similarity

Refence



(這裡紅框為Recall算法)

Candidate

Recall

$$R_{\mathrm{BERT}} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\boxed{\hat{x}_j \in \hat{x}}} \mathbf{x}_i^{\top}\hat{\mathbf{x}}_j$$

Based on reference

以 Reference tokens 的分數取最大值

Precision

$$P_{\mathrm{BERT}} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{\boxed{x_i \in x}} \mathbf{x}_i^{\top}\hat{\mathbf{x}}_j$$

Based on candidate

以 Candidate tokens 的分數取最大值

Importance Weighting (罕見字的影響)

我今天喝了一杯咖啡。
我今天喝了一杯抹茶拿鐵。
我今天去路易莎。

假設有一個字 w，以及全部有 M 篇文章

w 出現在 M 篇文章的次數為：$\sum_{i=1}^{M} \mathbb{I}[w \in x^{(i)}]$

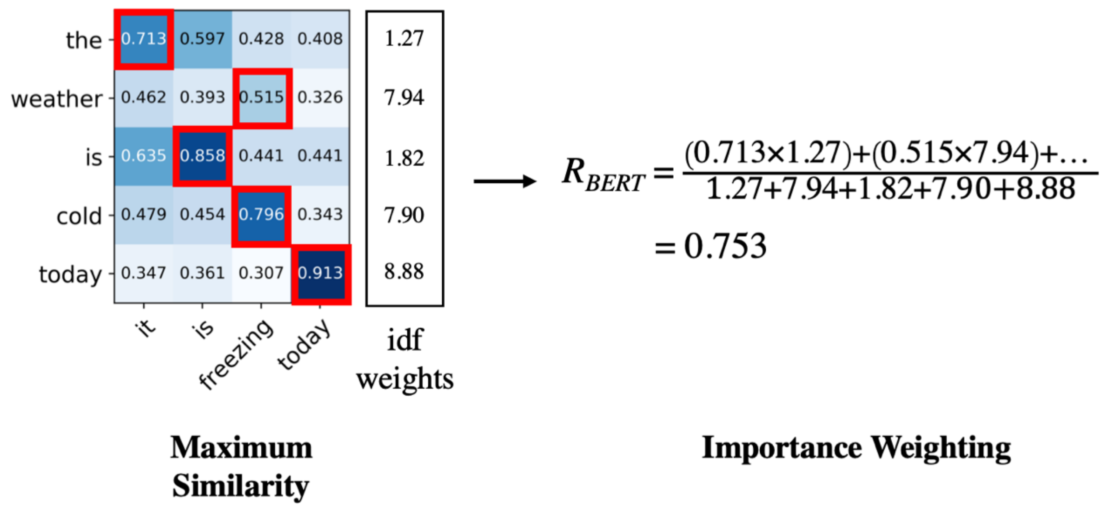
w 的 document frequency 為：$\frac{\sum_{i=1}^{M} \mathbb{I}[w \in x^{(i)}]}{M}$

w 的 inverse document frequency (IDF) 為：$\frac{M}{\sum_{i=1}^{M} \mathbb{I}[w \in x^{(i)}]}$

BERTScore – Importance Weighting

Given M reference sentences $\{x^{(i)}\}_{i=1}^{M}$, the idf (inverse document frequency) score of a word-piece token w is:

$$R_{\text{BERT}} = \frac{\sum_{x_i \in x} \text{idf}(x_i) \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j}{\sum_{x_i \in x} \text{idf}(x_i)}$$



Summary of BERTScore

- BERTScore leverages the contextual representation abilities of BERT to measure the semantic similarities between a reference and a candidate.
- In the paper, BERTScore correlates better with human judgments and provides stronger model selection performance than existing metrics.
- However, BERTScore does not involve training process.

Can we train BERT for a better evaluation metric?

Zhang, Tianyi, et al. "BERTScore: Evaluating Text Generation with BERT." International Conference on Learning Representations. 2020.

BLEURT – Quick Introduction

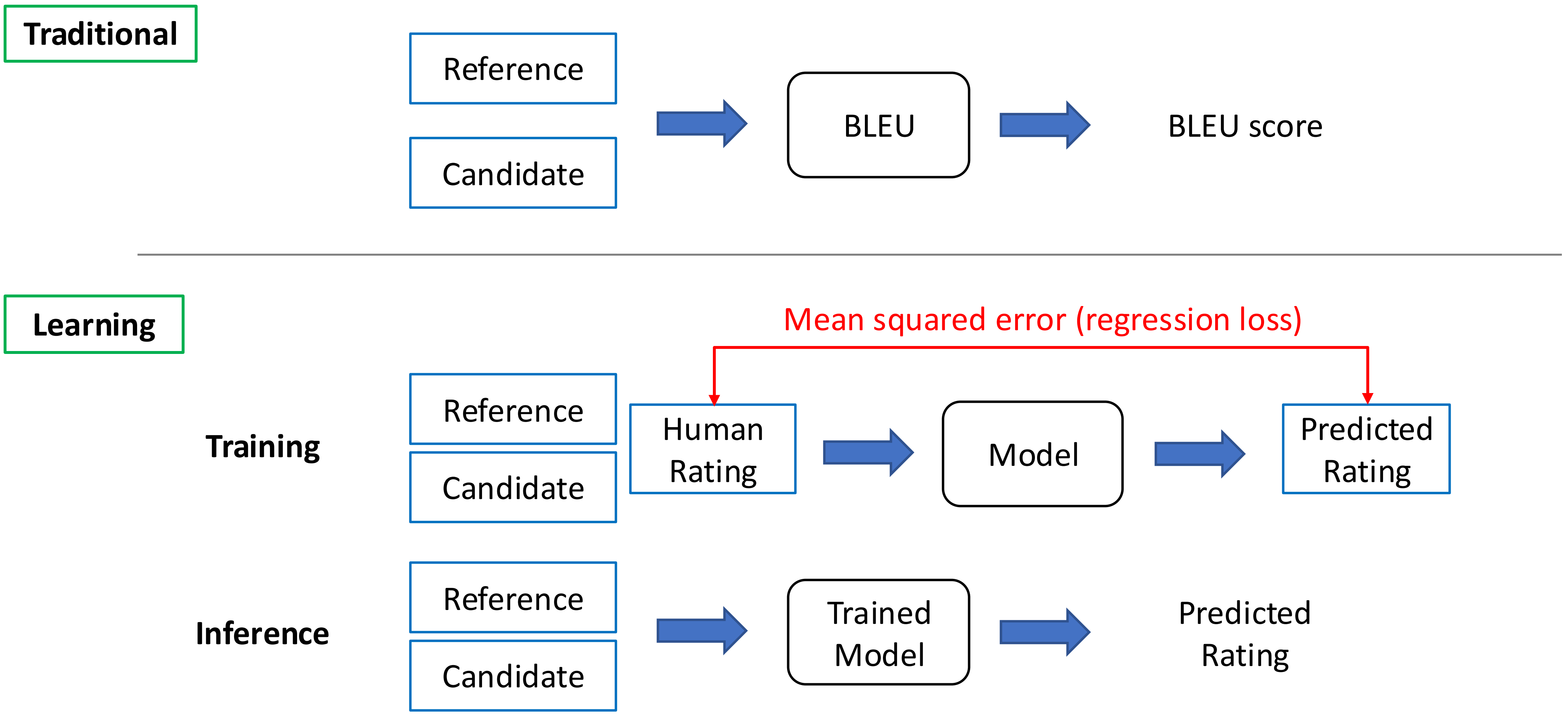
- BLEURT: Learning Robust Metrics for Text Generation, published by Google
- BLEURT trains BERT for a more robust evaluation metric.
  - Mainly for machine translation.
    - Also get hints from the name “BLEURT”
  - Trained checkpoint can be obtained. We don’t need to perform training.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. "BLEURT: Learning Robust Metrics for Text Generation." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

BLEURT – Motivations

- Learned metrics can be tuned to measure task-specific properties, such as fluency, faithfulness, grammar, or style.
- NLG systems tend to get better over time, and therefore a model trained on ratings data from 2015 may fail to distinguish top performing systems in 2019, especially for newer research tasks.

Training on Human Ratings



BLEURT 前情提要

- BLEURT 是一個 BERT (以英文BERT初始化)
- BLEURT 有 pre-training 跟 fine-tuning
  - fine-tuning: 學習人類的打分
  - pre-training: 使用 data augmentation 在非打分任務上面進行暖身 (warm-up, for transfer learning)

BLEURT – Steps

Step 0: Reference-candidate pairs $(z, \tilde{z})$ and the pre-trained BERT model

Step 1: Data augmentation for $(z, \tilde{z})$ to to perform pre-training

Data augmentation strategies

- Random masking
- Back-translation
- Dropping words randomly

Total 6.5 million variants of $(z, \tilde{z})$ were created.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. "BLEURT: Learning Robust Metrics for Text Generation." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

Random masking (for pre-training)

Two kinds of masking strategies were adopted:

Token masking

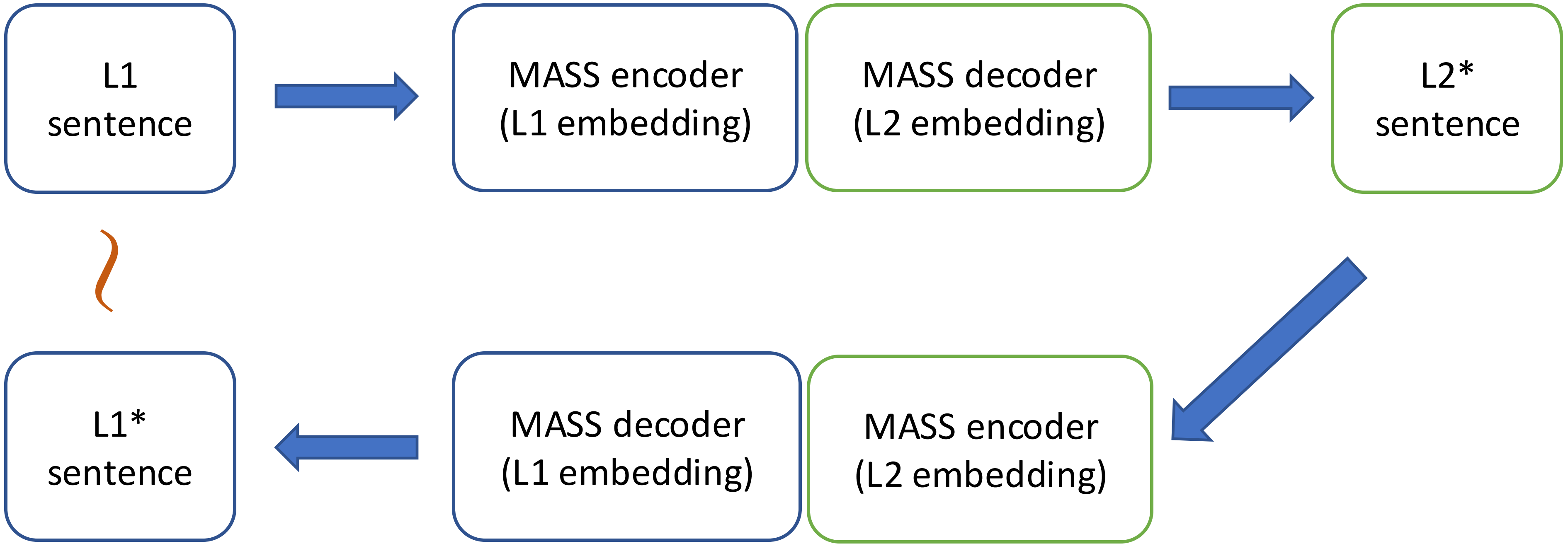
I love traveling to Vancouver for attending a conference.	→	I love traveling to Vancouver for **[MASK]** a conference.

Span masking

I love traveling to Vancouver for attending a conference.	→	I love traveling to Vancouver for **[MASK] [MASK] [MASK]**.

Backtranslation (for pre-training)

- For example: L1 -> English; L2 -> French or German



Dropping words randomly (for pre-training)

- The authors found it useful in their experiments to randomly drop words to create other examples.

I love traveling to Vancouver for attending a conference.	→	I love to Vancouver for attending a conference.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. "BLEURT: Learning Robust Metrics for Text Generation." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

BLEURT – Step 3 (for pre-training)

Step 3: Pre-training each sentence pair $(z, \tilde{z})$ with the following tasks.

Note that this is not conventional BERT pre-training! It is multi-task pre-training!

Task Type	Pre-training Signals	Loss Type
BLEU	$\boldsymbol{\tau}_{\text{BLEU}}$	Regression
ROUGE	$\boldsymbol{\tau}_{\text{ROUGE}} = (\tau_{\text{ROUGE-P}}, \tau_{\text{ROUGE-R}}, \tau_{\text{ROUGE-F}})$	Regression
BERTscore	$\boldsymbol{\tau}_{\text{BERTscore}} = (\tau_{\text{BERTscore-P}}, \tau_{\text{BERTscore-R}}, \tau_{\text{BERTscore-F}})$	Regression
Backtrans. likelihood	$\boldsymbol{\tau}_{\text{en-fr},\boldsymbol{z}\mid\tilde{\boldsymbol{z}}}, \boldsymbol{\tau}_{\text{en-fr},\tilde{\boldsymbol{z}}\mid\boldsymbol{z}}, \boldsymbol{\tau}_{\text{en-de},\boldsymbol{z}\mid\tilde{\boldsymbol{z}}}, \boldsymbol{\tau}_{\text{en-de},\tilde{\boldsymbol{z}}\mid\boldsymbol{z}}$	Regression
Entailment	$\boldsymbol{\tau}_{\text{entail}} = (\tau_{\text{Entail}}, \tau_{\text{Contradict}}, \tau_{\text{Neutral}})$	Multiclass
Backtrans. flag	$\boldsymbol{\tau}_{\text{backtran\_flag}}$	Multiclass

- Ground-truth values can be computed for each $(z, \tilde{z})$ pair!
- Losses for the six tasks were sum up during pre-training.

- Regression: mean squared error
- Multiclass: Cross-entropy

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. "BLEURT: Learning Robust Metrics for Text Generation." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

Task 4: Backtranslation Likelihood

- Existing FOUR translation models (trained) are needed.
  - Transformers (Vaswani et al., 2017): EN-FR, FR-EN
  - Transformers (Vaswani et al., 2017): EN-DE, DE-EN
- Equations use EN-FR for an example

$P(x_t|x_1, \dots, x_{t-1}, z)$

$$\boldsymbol{z}^*_{\texttt{fr}} = \arg\max P_{\texttt{en}\to\texttt{fr}}(\boldsymbol{z}_{\texttt{fr}}|\boldsymbol{z})$$

Best translated French sentence (details absent in the paper)

$$P(\tilde{\boldsymbol{z}}|\boldsymbol{z}) \approx P_{\texttt{fr}\to\texttt{en}}(\tilde{\boldsymbol{z}}|\boldsymbol{z}^*_{\texttt{fr}})$$

Backtranslation Likelihood

Task 4: Backtranslation Likelihood

先翻到法文，
再翻回英文

先翻到德文，
再翻回英文

Backtrans. likelihood $\mathcal{T}_{\text{en-fr},\boldsymbol{z}|\tilde{\boldsymbol{z}}}$, $\mathcal{T}_{\text{en-fr},\tilde{\boldsymbol{z}}|\boldsymbol{z}}$, $\mathcal{T}_{\text{en-de},\boldsymbol{z}|\tilde{\boldsymbol{z}}}$, $\mathcal{T}_{\text{en-de},\tilde{\boldsymbol{z}}|\boldsymbol{z}}$

使用第二組的結果
先翻到法文
再翻回英文原句

使用第四組的結果
先翻到德文
再翻回英文原句

符號提醒：z跟波浪z都是同語言

Task 5 and Task 6

- **Textual Entailment**
  - We report the probability of three labels: Entail, Contradict, and Neutral, using BERT fine-tuned on the MNLI dataset.
- **Backtranslation flag**
  - A Boolean that indicates whether the perturbation was generated with backtranslation or with mask-filling (例如替換 tokens).

BLEURT – Final Step

Step 4: Fine-tune the model (trained from Step 3) on the <Reference, Candidate, Rating> data using the regression loss (Mean squared error)

The <Reference, Candidate, Rating> data include

- WMT (machine translation task)
- WebNLG (for general text generation)
    - **semantics**, **grammar**, and **fluency**

Summary of BLEURT

- This approach uses (continual) pre-training and fine-tuning to create a learned evaluation metric for machine translation and general NLG.
- According to the paper, BLEURT is better aligned to human ratings then BERTScore.
- BLEURT should work for text summarization, but the authors did not test it.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. "BLEURT: Learning Robust Metrics for Text Generation." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

Comparison for Human and Automatic Evaluations (e.g., BLEU and ROUGE)

- Human evaluations
  - Pros: More accurate for subjectivity, flexibility for any desired comparison
  - Cons: Less objective, time-consuming, expensive
- Automatic evaluations
  - Pros: Objective enough to serve as common evaluation metrics, fast
  - Cons: Cannot meet language diversity
    - Take machine translation for instance, there are always other valid ways to translate the source sentence.

GPTRank

https://aclanthology.org/2024.naacl-long.478

You will be given a news article along with two summaries. Please compare the quality of these two summaries and pick the one that is better (there can be a tie). First you will give an explanation of your decision then you will provide your decision in the format of 1 or 2 or tie.

Response format:
Explanation: "Your explanation here".
Decision: 1 or 2 or tie.
Here's the article:
{{Article}}
Summary 1:
{{Summary 1}}
Summary 2:
{{Summary 2}}

...

Thank you!

Instructor: 林英嘉

✉ yjlin@cgu.edu.tw