Writing Homework 1 – 309833027 樂又瑄

1. BLEURT

(1) 一般使用人工和自動化評估 NLG 模型，例如 BLEU 或 ROUGE，但是 BLEU 或 ROUGE 無法識別句子語意與語法變化的問題。因此 Google 發展一種以 BERT 為基礎、模擬人工判斷的評估方式，具備人工和自動評估的優點。

因訓練資料相當有限，對於 ML 模型訓練用的資料相當不足。因此應用一種用於語意理解的非監督式模型 BERT 的上下文單詞方法。而為了提高 BLEURT 的 robustness，Google 發展了一種新穎的 pre-train 方法：在人工評估微調前，使用大量來自維基百科加上隨機擾動生成的合成句對模型進行「預熱」。

(2) Pros and Cons : BLEURT 的優點為可以捕捉句子之間語意相似性，準確率比 BLUE 高約 48%；缺點為 BLEURT 的評分方式較保守，不一定能給予 NLG 的性能較好的評分。

2. UNION

(1) 現有的參考指標(BLEU 和 MoverScore)對於開放式 text generation 的判斷相關性較差，對於相同的輸入詞，可能在語義上有所不同。因此提出以 BERT 為基礎的 UNION，可用於開放式 story generation。訓練時不需要特定的 NLG 模型或人工註釋，且使用四種負樣本解決生成故事中常見的問題。

(2) Pros and Cons : UNION 的優點為可以不需要借助人工註釋或 NLG 模型，可以更快速的評估故事生成好壞