

## 基於訊息配對相似度估計的聊天記錄解構

劉至咸 ZhiXian Liu

國立中央大學資訊工程學系

Department of Computer Science and Information Engineering

National Central University

zhixian989@gmail.com

張嘉惠 Chia-Hui Chang

國立中央大學資訊工程學系

Department of Computer Science and Information Engineering

National Central University

chia@csie.ncu.edu.tw

### 摘要

一般而言，為建立 Retrieval-based 聊天機器人，我們可以從聊天紀錄中來產所需的問答配對 (Question-Answer Pair)，然而問答配對並非完全依序地呈現在聊天紀錄中，不同內容的問答配對可能互相穿插，而從互相穿插的訊息中分離出內容不同的會話的任務稱為對話解構 (conversation disentanglement)。

現有的對話解構研究大多透過計算兩個訊息的相似度來解決問題，在此論文中，我們發透過計算訊息相似度判斷訊息是否屬於相同會話是非常困難的，但若我們透過計算相似度來預測訊息的回覆關係則可以解決此問題。此外我們指出過去研究中的模型無法處理未經訓練的訊息，並無法在實務上運用的缺陷。

此論文中，我們使用 IRC 與 Reddit 資料集進行實驗，並使用聊天記錄進行對話解構。其中人工合成的 Reddit 資料集提供額外的大量訓練資料，且 BERT 模型在此資料集上的回覆關係預測獲得良好的效能。

關鍵詞：對話解構，回覆關係預測，BERT 模型應用