Computational Linguistics and Chinese Language Processing
Vol. 24, No. 2, December 2019, pp. 19-36

EBSUM: 基於 BERT 的強健性抽取式摘要法

EBSUM: An Enhanced BERT-based Extractive Summarization Framework

吳政育*、陳冠宇*

Zheng-Yu Wu and Kuan-Yu Chen

摘要

目前大部分自動摘要方法，分為抽取式摘要(Extractive)與重寫式摘要(Abstractive)，重寫式摘要雖然能夠改寫文章形成摘要，但這並不是一種有效的方式，困難點在於語意不通順、重複字等。抽取式摘要則是從文章中抽取句子形成摘要，能夠避免掉語意不通順，重複字的缺點。目前基於BERT(Bidirectional Encoder Representation from Transformers)的抽取式摘要法，多半是利用 BERT 取得句子表示法後，再微調模型進行摘要句子之選取。在本文中，我們提出一套新穎的基於 BERT 之強健性抽取式摘要法(Enhanced BERT-based Extractive Summarization Framework, EBSUM)，它不僅考慮了句子的位置資訊、利用強化學習增強摘要模型與評估標準的關聯性，更直接的將最大邊緣相關性(Maximal Marginal Relevance, MMR)概念融入摘要模型之中，以避免冗餘資訊的選取。在實驗中，EBSUM 在公認的摘要資料集 CNN/DailyMail 中，獲得相當優良的任務成效，與經典的各式基於類神經網路的摘要模型相比，EBSUM 同樣可以獲得最佳的摘要結果。

Abstract

Automatic summarization methods can be categorized into two major streams: the extractive summarization and the abstractive summarization. Although abstractive summarization is to generate a short paragraph for expressing the original document, but most of the generated summaries are hard to read. On the contrary, extractive summarization task is to extract sentences from the given document to

* Departmet of Computer Science and Information Engineering, National Taiwan University of Science and Technology, CSIE
E-mail: {M10615079; kychen}@mail.ntust.edu.tw

construct a summary. Recently, BERT (Bidirectional encoder representation from transformers), which has been introduced to several NLP-related tasks and achieved remarkable results, is a pre-trained language representation method. In the context of extractive summarization, BERT is usually be used to obtain representations for sentences and documents, and then a simple model is employed to select potential summary sentences based on the inferred representations. In this paper, an enhanced BERT-based extractive summarization framework (EBSUM) is proposed. The major innovations are: first, EBSUM takes the sentence position information into account; second, in order to maximize the ROUGE score, the model is trained by the reinforcement learning strategy; third, to avoid the redundancy information, the maximal marginal relevance (MMR) criterion is incorporated with the proposed EBSUM model. In the experiments, EBSUM can outperforms several state-of-the-art models on the CNN/DailyMail corpus.

關鍵詞：自動摘要，抽取式，BERT，強化學習，最大邊緣相關性
Keywords: Auto-summarization, Extractive, BERT, Reinforcement Learning, MMR

1. 緒論 (Introduction)

隨著網際網路的普及，傳統的電子佈告欄系統與新興的社群媒體、數位討論平台日益蓬勃，大量的文字資訊充斥在網路媒體上。有趣的是，雖然存在大量可利用的文字資料，但同時卻存在資訊氾濫的問題。資料檢索(Information Retrieval)的任務是依據使用者輸入的查詢(Query)，對所有文章進行排序，從大量的資訊中擷取使用者需要的文章(Nogueira & Cho, 2019; Wu, Yen, & Chen, 2019; Yang, Zhang, & Lin, 2019)。雖然資訊檢索已能篩選出使用者最可能需要的文章，但文章數量仍然相當龐大、內容繁雜且冗餘，使用者需要耗費多餘的時間處理不必要的資訊，以便得知各篇文章的有意義的部分。為了進一步地幫助使用者處理大量的資料(Pak & Paroubek, 2010)，自動文件摘要技術(F. Liu, Flanigan, Thomson, Sadeh, & Smith, 2018; Rambaut, Drummond, Xie, Baele, & Suchard, 2018)成了一個重要的研究議題，他不僅可以快速地整理出每一篇文章的重點資訊，讓使用者篩選自己所需要的資訊，在行動裝置的普及的世代中，如何再低頻寬下將資訊整理，在有限的顯示面積下，呈現給使用者更多的資訊，也成了自動摘要一個很好的應用場景(Billawala, Mehdad, Radev, Stent, & Thadani, 2018; Leiva, 2018)。除此之外，電子媒體的發展，為了商業行銷、剖析民論輿情或是調查社會意向等，促使自然語言處理與分析成為重要的研究領域。其中，摘要的抽取與分析，扮演著很重要的角色。因此，文件摘要不僅常被用於傳統的新聞大綱任務中(Maskey & Hirschberg, 2003; Maybury & Merlino Jr, 2005)，亦被應用於輿情分析的任務中(Fan & Gordon, 2014; Imran, Castillo, Diaz, & Vieweg, 2015; Stieglitz & Dang-Xuan, 2013)。輿情系統必須對文章進行情緒分析判斷其為正面或反面意見(Fan & Gordon, 2014; He, Wu, Yan, Akula, & Shen, 2015)，也需讓使用者更快速瞭解其

所想知道的正反意見文章的內容，所以自動文件摘要就成為輿情系統中重要的一部分。

摘要的概念為使用較少的文字表達出文章中重要的中心意義，讓人們可以快速的篩選自己所需要的資訊，減少吸收資訊的時間，但卻能了解到原始文章的內容，因此文件自動摘要能有效解決資訊氾濫的問題。文件摘要最開始是由人工方式來進行文章摘要的作業，雖然這樣的人工摘要品質高，但遇到大量且更新快速的應用，例如：線上新聞網站、線上論壇、電子郵件…等等，必須耗費大量金錢以及時間，若自動摘要能夠有效套用至上述各個應用，可為企業以及使用者提供更快速的閱讀體驗。然而，在現階段的應用上，最常見且堪用的自動摘要方法是領導摘要(Lead)(See, Liu, & Manning, 2017)，領導摘要是以句子作為基本單位來建構的，也就是將文章前幾句的句子蒐集並形成一個摘要。雖然此類利用規則所建立之摘要有一定的效果、且易於實現，但其缺點是只能產生堪用的摘要，並且只有像是新聞文章，通常以破題的方式在前幾句概述本篇新聞的重點資訊，而適合用領導摘要法產生摘要。除了在文章中選取句子產生摘要的抽取式摘要法(Extractive Summarization)外，重寫式摘要(Abstractive Summarization)也是目前主流的研究方向。重寫式摘要需要電腦先對文章進行閱讀與理解，接著以生成的方式重寫出文章的摘要。因此，重寫式摘要似乎可以產生較自然的摘要內容，但也可能因為產生不通順的語句，而降低閱讀的流暢性(F. Liu *et al*., 2018; L. Liu *et al*., 2018; Nallapati, Zhou, dos Santos Gulcehre, & Xiang, 2016; See *et al*., 2017)；抽取式摘要是由文章中選取完整的句子作為摘要，因此摘要的可讀性通常是較佳的，也因此，抽取式摘要法仍然是許多學者研究的議題(Nallapati, Zhai, & Zhou, 2017; Narayan, Cohen & Lapate, 2018; Wong, Wu, & Li, 2008)。

近年來，由於 BERT 模型(Devlin, Chang, Lee, & Toutanova, 2018)的提出，許多自然語言處理的任務皆取得了突破性的進展。在抽取式摘要的研究中，BERTSUM(Y. Liu, 2019)使用 BERT 取得每個句子的表示法，然後利用後續的分類器，為每一個句子評分，作為是否選取句子的依據。此外，為了減少摘要的冗餘，BERTSUM 採用規則式的三連詞過濾法(Trigram Block)，捨棄與已選摘要存在重複三連詞的候選句子，藉此減少摘要中冗餘的資訊。雖然 BERTSUM 已在抽取式摘要任務中取得相當優良的任務成效，但我們認為，BERTSUM 缺乏考量句子在文章中的位置資訊；此外，BERTSUM 是針對每一個句子以計算交叉熵(Cross Entropy)進行訓練，目的為最大化與正確摘要句子的似然值(Likelihood)，但交叉熵優化方式不需要對句子進行排名，且與摘要的評分方法 ROUGE 之間不存在對應的關係；還有，BERTSUM 僅使用簡單的三連詞過濾方法，減少冗餘資訊的選取。有鑑於此，本論文提出一套基於 BERT 的強健性摘要方法 EBSUM(Enhanced BERT-based Extractive Summarization Framework)，他不僅考量了句子在文章中的位置資訊，利用強化學習(Narayan *et al.*, 2018)的方式讓摘要模型與評估方式的關係更為緊密，也進一步地讓摘要模型本身具備減少冗餘資訊選取的能力。

2. 預訓練之詞向量表示法 (Pretained Word Embedding)

分布式向量表示法又稱詞向量(Word Embeddings)，是在自然語言處理中被廣泛使用的方法，其目的是將每一個單詞以一個低維空間的分布式向量表示之。早期經典的方法有連續詞袋模型(Continuous Bag-of-words, CBOW)、跳字模型(Skip-gram) (Mikolov, Chen, Corrado, & Dean, 2013)以及全局向量 (Global Vectors, GloVe) (Pennington, Socher, & Manning, 2014)。在這些經典的模型中，每一個字詞在不同的上下文中是以相同的詞向量表示之，但許多字詞在不同的上下文中有著不同的含意，例如“蘋果”一詞，依照其上下文，可能代表著手機品牌或者是一種水果。因此，為了讓每一個詞的表示法更加強健，2018 年 Peter 首先提出 ELMo(Peters *et al*., 2018)架構，利用雙向長短期記憶模型(BiLSTM)考慮上下文的資訊，為每一個位置的詞輸出相對應的詞向量，也就是在不同位置的同一個詞，將有著不同的低維度向量表示法。接著，由於長短期記憶模型(Long Short-term Memory, LSTM)在模型參數更新時，無法以平行運算的方式加速計算，並且在一些任務上已被證明 Transformer(Vaswani *et al*., 2017)架構的優點與效能，因此 OpenAI 基於 Transformer 架構，提出 GPT(Generative Pre-training) (Radford, Narasimhan, Salimans, & Sutskever, 2018)模型，用於學習每一個詞的詞向量表示法。隨後，基於 GPT，谷哥又再提出 BERT(Bidirectional Encoder Representations from Transformers) (Devlin *et al*., 2018)。

相較於 GPT，BERT 在模型架構上進行些許變動，首先利用 Transformer 中的 Encoder 取代 GPT 中的 Transformer Decoder，並採用雙向（Bidirectional）語言模型方式進行預訓練，且訓練資料比 GPT 更大。此外，BERT 屬於多任務學習，在模型訓練時分為兩個步驟，第一步驟是預訓練遮罩語言模型(Masked Language Model)，即隨機遮罩訓練資料中百分之十五的單字量，並用[MASK]代替，讓模型利用[MASK]的上下文資訊預測 [MASK]原本的正確單詞：

$$今天天氣真[MASK]，適合去[MASK]野餐 \tag{1}$$

為了防止模型無法收斂，在實作過程中，80%的訓練時間是使用[MASK]遮罩單詞，另外 10%的訓練時間則使用一個隨機的詞當作遮罩，以及 10%的訓練時間給定正確的詞。第二步驟，為了使 BERT 考慮到句子等級關係，因此除了遮罩語言模型，加入句子關聯性的分類任務(Next Sentence Prediction)，即給予兩個句子，判斷第二句是否為第一句的後續句子。更明確地，訓練 BERT 模型時的輸入如圖 1 所示，[CLS]表示句子的開頭，而

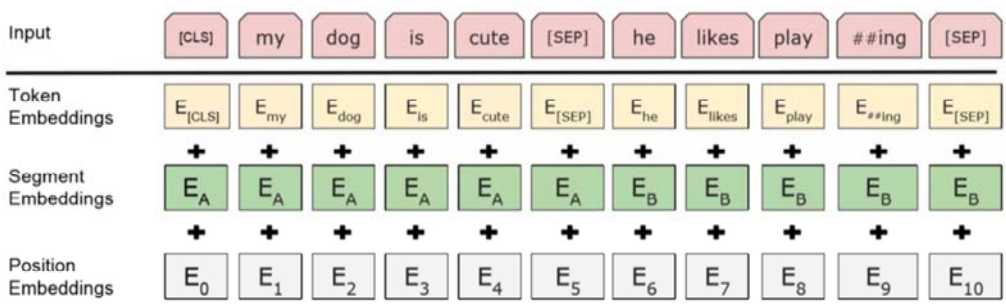

圖 1. BERT 架構圖。
[Fiqure 1. Illustration of BERT Model.]

[SEP]表示分句，且在預訓練中每個位置的字詞由三種向量表示法相加而成，分別是位置向量(Position Embedding)、詞向量(Token Embedding)以及段落向量(Segment Embedding)。位置向量是用來表示這個字詞是位於輸入序列中的哪一個位置；段落向量則用於表示單詞是位於上句 E_A 或者下句 E_B；最後，BERT 的輸入為此三種向量相加形成，而[CLS]可以視為是整體的表示法，並且被用於句子關聯性的分類任務之中。

3. 基於BERT的抽取式摘要方法 (BERT-Based Extractive Summarization Method)

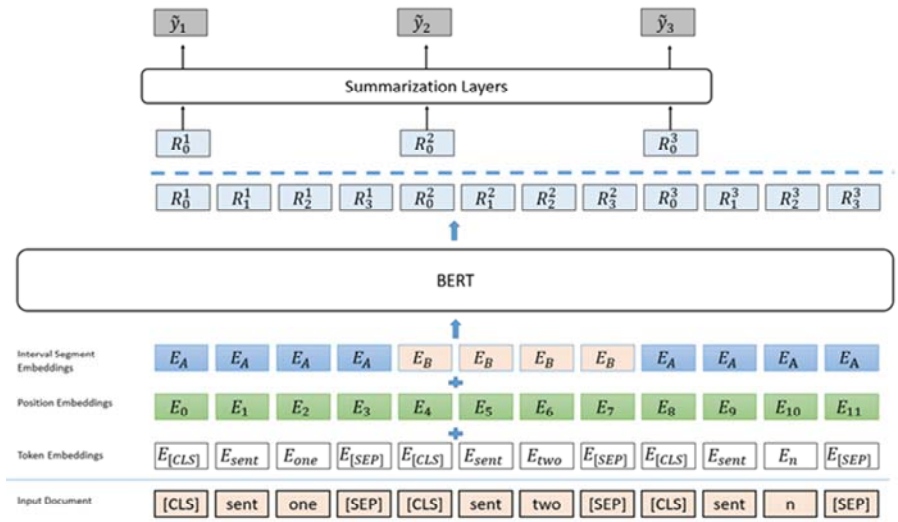

圖 2. 基於 BERT 的基準抽取式摘要法(BERTSUM(Y. Liu, 2019))架構圖。
[Fiqure 2. Illustration of BERTSUM Model.]

3.1 BERTSUM

由於 BERT 幾乎在所有類型的自然語言處理任務皆得到了相當大幅的效能提升，包括問答系統(Question Answering) (Qu *et al*., 2019; Yang, Xie *et al*., 2019)、資訊檢索(Wu *et al*., 2019; Yang, Zhang *et al*., 2019)以及對話系統(Dialog System)。在自動摘要(Gu & Hu, 2019; Lebanoff *et al*., 2019; Y. Liu, 2019; Zhang *et al*., 2019; Zheng & Lapata, 2019b)的研究中，BERT 也已被用於抽取式摘要的任務之中。BERTSUM (Y. Liu, 2019)將一篇給定的文章D視為一連串句子的集合：

$$D = [sent_1, sent_2, \cdots, sent_n] \tag{2}$$

其中，$sent_i$表示在D中第i個句子。在輸入的階段，BERTSUM 在每一個句子的開頭加入[CLS]，並且在句子結尾加入[SEP]；另外，與 BERT 原論文相同，每一個句子皆有一個段落向量；最後，每一個字的向量表示法即是將位置向量、詞向量以及段落向量相加而

成。BERTSUM 的特點是將第i個[CLS]標籤在 BERT 中最後一層的輸出，當作$sent_i$的表示法R_i。在訓練抽取式摘要器時，BERTSUM 將每一個句子$sent_i$由 BERT 所求得的表示法R_i經過分類器輸出分數，判斷是否為正確的摘要句子。若為正確的摘要句子，其值應為 1；反之，不是摘要句子的分數應為 0：

$$\begin{cases} \tilde{y}_i = 0 & if\ sent_i\ not\ in\ summary \\ \tilde{y}_i = 1 & if\ sent_i\ in\ summary \end{cases} \tag{3}$$

BERTSUM 提出的三種分類器，包括簡單分類器(Simple Classifier)、Transformer 以及遞迴神經網路(Recurrent Neural Network, RNN) (Hochreiter & Schmidhuber, 1997)，並且利用二分類交叉熵(Binary Cross Entropy)計算預測誤差，並依此為整個 BERTSUM 模型進行參數的訓練與調整。BERTSUM 的模型架構如圖 2 所示。值得注意的是，在測試階段(Test Stage)，為了增加摘要內容的多樣性、避免選取出冗餘的句子，BEETSUM 使用三連詞過濾方法(Carbonell & Goldstein, 1998)，以減少摘要的冗餘，當給定已選摘要句子集$\check{S}$以及 BERTSUM 所預測出的候選摘要句子c，若c與$\check{S}$存在任一三字元組，則忽略c，使得摘要中每一個句子不相互重疊。

3.2 PACSUM (Zheng & Lapata, 2019a)

BERTSUM 在抽取式摘要任務中取得相當優良的任務成效，但是監督式學習方式必須依賴大量的標記訓練集，不易應用於現實應用中(Cheng & Lapata, 2016; Gehrmann, Deng, & Rush, 2018; Nallapati *et al*., 2017; Nallapati *et al*., 2016; Narayan *et al*., 2018; Paulus, Xiong, & Socher, 2017; See *et al*., 2017)，因此無監督學習仍然是許多任務的研究方向(Erkan & Radev, 2004; Hirao, Yoshida, Nishino, Yasuda, & Nagata, 2013; Li, Wang, Lam, Ren, & Bing, 2017; Lin & Hovy, 2002; Marc, 1998; Mihalcea & Tarau, 2004; Parveen, Ramsl, & Strube, 2015; Radev, Jing & Budzikowska, 2000; Wan, 2008; Wan & Yang, 2008; Yin & Pei, 2015)。大部分無監督式學習的抽取式摘要任務採用基於圖(Graph-based)的排序演算法，用以計算句子在文章中的顯著性(Salience)。更明確地，當將給定文章$D = [sent_1, sent_2, \cdots, sent_n]$，可將每個句子$sent_i$表示為圖中的節點(node)，而任二個節點$(sent_i, sent_j)$間有邊$e_{ij}$相連，並且以相似度分數(similarity)作為邊e_{ij}的權重。接著，再以各式圖論的演算法，如 Pagerank (Brin & Page, 1998)，計算每一句子的中心性(centrality)：

$$\text{centrality}(sent_i) = \sum_{j\in\{1,\ldots,i-1,i+1,\ldots,n\}} e_{ij} \tag{4}$$

此一中心性分數就被用來做為該句子是否為摘要句子的分數。

PACSUM 認為 Pagerank 並無考慮到圖的方向性(Undirected)，因此 PACSUM 透過相對位置來計算中心性，希望越早出現的句子越重要。基於這個想法，PACSUM 計算$sent_i$中心性分數的方式為：

$$\text{centrality}(sent_i) = \lambda_1 \sum_{j<\text{i}} e_{ij} + \lambda_2 \sum_{j>\text{i}} e_{ij} \tag{5}$$

其中，λ_1以及λ_2分別為前向(Forward-looking)以及後向(Backward-looking)的權重，且

$\lambda_1 + \lambda_2 = 1$，在實驗中 PACSUM 發現$\lambda_1 < 0$能夠有最佳的效果，這表明了若句子與前面的句子相似的話，會使中心性分數降低。除了分數計算的方式改變外，PACSUM 使用 BERT 做為句子的編碼器，並以句子層級的分佈假設(Sentence-level distributional hypothesis) (Harris, 1954; Polajnar, Rimell, & Clark, 2015)進行 BERT 參數的微調：

$$log\sigma\left(v'_{sent_{i-1}}{}^{T} v_{sent_i}\right) + log\sigma\left(v'_{sent_{i+1}}{}^{T} v_{sent_i}\right) + \mathbb{E}_{sent_i \sim P(s)}\left[log\sigma\left(-v'_{sent_i}{}^{T} v_{sent_i}\right)\right] \quad (6)$$

其中v'_{sent}以及v_{sent}為兩個不同的 BERT，σ為激活函數 sigmoid，$P(s)$為均勻分佈的句子空間。模型訓練完成後，可用於產生句子的向量表示法。

4. EBSUM：基於BERT的強健性抽取式摘要方法 (Enhanced BERT-based Extractive Summarization Framework, EBSUM)

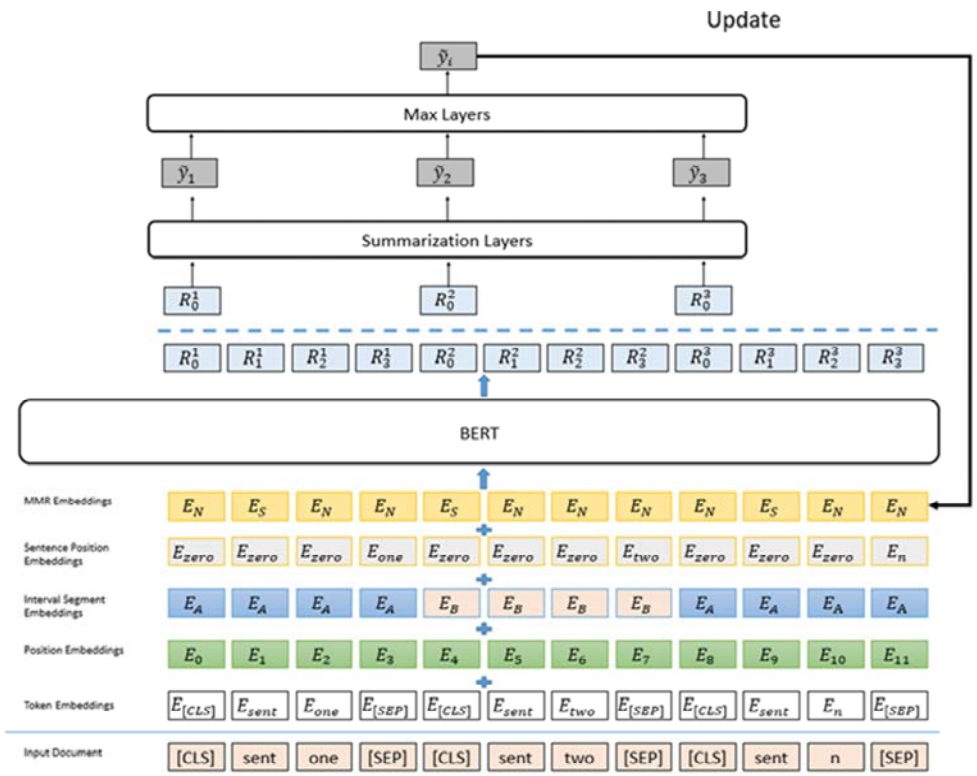

圖 3. 基於 BERT 的強健性抽取式摘要法(EBSUM)架構圖。
[Fiqure 3. Illustration of EBSUM Model.]

雖然 BERTSUM 已在抽取式摘要任務中取得相當優良的任務成效，但我們認為，BERTSUM 缺乏考量句子在文章中的位置資訊；此外，BERTSUM 是針對每一個句子以計算交叉熵(Cross Entropy)進行訓練，目的為最大化與正確摘要句子的似然(Likelihood)，

但交叉熵優化方式不需要對句子進行排名，且與摘要的評分方法 ROUGE 之間不存在對應的關係；還有，BERTSUM 使用三連詞過濾方法，減少冗餘資訊的選取。有鑑於此，本論文提出一套基於 BERT 的進階版摘要方法 EBSUM(Enhanced BERT-based Extractive Summarization Framework)，他不僅考量了句子在文章中的位置資訊，利用強化學習的方式讓摘要模型與評估方式的關係更為緊密，也進一步地讓摘要模型本身具備減少冗餘資訊選取的能力。

我們所提出之 EBSUM 摘要模型架構如圖 3 所示。首先，在輸入的階段，EBSUM 將文章D視為一連串句子的集合$D = [sent_1, sent_2, \cdots, sent_n]$，$sent_i$表示文章$D$中第$i$個句子，並且在每一個句子的開頭加入[CLS]、句子結尾加入[SEP]。

4.1 句子位置向量 (Sentence Position Embedding)

許多研究指出人們在撰寫文章時，會習慣在文章前半部講述重點，尤其新聞文章更是常見，因此有許多自動摘要模型，會將句子在文章中的位置資訊作為一項特徵。為了將此一特徵納入運用，EBSUM 提出三種考慮句子位置資訊的特徵：CLS 向量、ALL 向量以及 SEP 向量，如圖 4 所示。CLS 向量是在每一句$sent_i$的開頭符號[CLS]位置上，加上對應該句的位置向量E_i，其中i為句子$sent_i$在文章的位置，而除了[CLS]以外的字詞，通通加上一個相同的特徵E_{zero}；ALL 向量則是將每一個句子$sent_i$中的每一個字詞都加入代表這一個句子的位置向量E_i；與 CLS 向量相反，SEP 向量是在每一句$sent_i$的結尾符號 [SEP] 位置上，加上對應該句的位置向量E_i，其中i為句子$sent_i$在文章的位置，而除了[SEP]以外的字詞，通通加上一個相同的特徵E_{zero}。

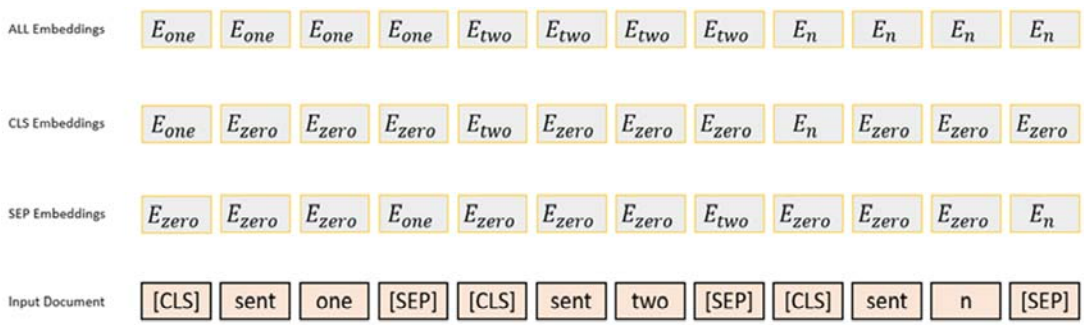

圖 4. 三種句子位置向量，分別為 SEP 向量、CLS 向量、ALL 向量。
[Fiqure 4. SEP Embeddings 、CLS Embeddings and ALL Embeddings.]

4.2 強化學習(Reinforcement Learning)

在抽取式摘要任務中，許多模型利用神經網路對文章中的每一個句子進行各式特徵抽取，藉由這些特徵，預測每個句子被選為摘要句的分數，並使用交叉熵(Cross Entropy)計算當前模型的預測結果與正確答案的差異，再進行模型參數的更新。這樣的做法是希望正確

的摘要句子可以獲得較高的預測似然值(Likelihood)，但以優化每一個句子的交叉熵值為目標，會忽略句子與句子之間的排名關係，並且這樣的訓練方式並無法直接與摘要的分數相對應。所以，有研究指出，使用交叉熵作為優化的目標容易產生過長的摘要或者選取到冗餘的句子(Narayan *et al.*, 2018)。有鑑於此，本研究透過強化學習(Sutton & Barto, 2018)中的策略學習(Policy Learning) (Williams, 1992)，希望讓模型的訓練目標與摘要分數（即 ROUGE）有更明確的對應關係。在強化學習中，對於給定的一篇文章D，我們可以隨機的抽取一組句子S形成摘要，並以 ROUGE-1 與 ROUGE-2 的 F1 分數總和做為這組句子對應的獎勵值$r(S)$。因此，強化學習的目標函式為最小化預期負面獎勵值(Negative Expected Reward)：

$$L(\Theta) = -\mathbb{E}_{S\sim P(\cdot|D,\Theta)}[r(S)] \tag{7}$$

其中Θ是抽取式摘要模型參數。當使用強化學習於抽取式摘要的任務上時，實作上我們在其交叉熵損失上與$r(S)$相乘，其目標是讓模型懂得區別哪些句子容易出現在高獎勵值的組合中，哪些句子則通常會讓獎勵值變低，因此達到抽取式摘要模型懂得如何抽取適當的句子組成摘要。

4.3 最大邊緣相關性向量 (Maximal Marginal Relevance, MMR)

為了使抽取式摘要模型可以自動地考量冗餘資訊的問題，我們發想於資訊檢索領域中的最大邊緣相關性(Maximal Marginal Relevance, MMR)準則，提出最大邊緣相關性向量，使得 EBSUM 在每個世代(iteration)都可以選擇出新穎且富有資訊的句子做為摘要。更明確地，當給定摘要文章$D = [sent_1, sent_2, \cdots, sent_n]$以及已選取的摘要集$\check{S}$，對於文章中已出現在$\check{S}$內的字詞，我們給予一個最大邊緣相關性向量$E_S$；反之，文章中未在$\check{S}$中出現的字詞，以一個最大邊緣相關性向量$E_N$表示之，其初始值為隨機向量，並交由 BERT 訓練學習，如圖 5 所示。為了讓 EBSUM 確實了解最大邊緣相關性向量的意義，在訓練時，我們將訓練資料中的每一篇文章隨機的挑選 0~3 句正確摘要句子放入集合$\check{S}$中，模型參數的更新目標則是正確地選取剩餘的正確摘要句子。

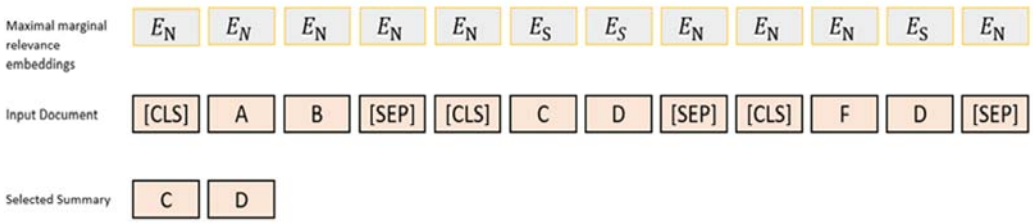

圖 5. 最大邊緣相關性向量。
[Fiqure 5. Maximal Marginal Relevance.]

4.4 微調抽取式摘要任務之模型 (Fine-tuned Extractive Summarization Model)

為了讓 EBSUM 考量句子在文章中的位置資訊以及具備抵抗冗餘資訊的能力，因此除了位置向量、詞向量以及段落向量外，我們額外引入了句子位置向量以及最大邊緣相關性向量，所以在 EBSUM 中，每一個字的向量表示法即是將這些向量相加而成。接著，藉由 BERT，我們可以得到每一個句子的向量表示法，即第i個[CLS]標籤在 BERT 中最後一層的輸出，當作$sent_i$的表示法R_i。在獲得每一個句子的表示法後，我們將這些向量輸入由多層 Transformer 堆疊而成的摘要任務(Summarization Layer)層(Ba, Kiros, & Hinton, 2016; Vaswani *et al*., 2017)，最終輸出每個句子被選擇為摘要的機率。值得一提的是，由於最大邊緣相關性向量的加入，EBSUM 是遞迴式的每次選取該次預測分數最大的句子加入已選摘要集Ś中，直到句子數（或預先設定的摘要比例或字數）達到預先設定的數量後，整個過程即停止。

5. 實驗結果與討論 (Experiment and Discussion)

我們使用 CNN/DailyMail (Hermann *et al*., 2015; See *et al*., 2017)資料集進行摘要模型的效能評估，含新聞文章和其摘要文章，平均字數以及句數如表 2，其中訓練集、驗證集和測試集分別有 287,227、13,368 以及 11,490 篇文章。我們使用 PyTorch、OpenNMT(Klein, Kim, Deng, Senellart, & Rush, 2017)以及 bert-base-uncased 版本的 BERT 實現本研究所提出之 EBSUM。為了進行比較，我們亦重現了強健的基準系統 BERTSUM。值得一提的是，在BERTSUM的前處理中，會將少於5個字詞的句子進行移除，可能是因為BERTSUM認為少於 5 個字的句子資訊量不足，但為了與其他基準系統進行比較，我們亦實作將少於 5 個字詞的句子進行保留，其餘設定不變之 BERTSUM 與領導摘要法(LEAD)，實驗結果如表 1 所示。這組實驗呈現了本研究所實現的 BERTSUM 與原論文的結果非常接近，也因此當我們將僅含有 5 個字以下的句子保留的實驗結果，可公平地與其他方法互相比較。

表 1. LEAD 與 BERTSUM 在 CNN/DailyMail 資料集上的實驗結果。
[Table 1. Experimental Results of the LEAD and BERTSUM in CNN/DailyMail]

		ROUGE-1	ROUGE-2	ROUGE-L
移除 5 個字以下的句子	LEAD (Y. Liu, 2019)	40.42	17.62	36.67
	LEAD	40.42	17.62	36.66
	BERTSUM (Y. Liu, 2019)	43.25	20.24	39.63
	BERTSUM	43.15	20.16	39.56
保留 5 個字以下的句子	LEAD	40.32	17.56	36.58
	BERTSUM	43.25	20.20	39.61

表 2. CNN/DailyMail 資料集的統計數據，分別以單詞與句數計算平均文章以及摘要長度。
[Table 2. Avg words and sentences of Documents and Summary in CNN/DailyMail]

		文章平均		摘要平均	
	總文章數	字數	句數	字數	句數
CNN+DM	11490	691.9	28.0	54.6	3.9

在第二組實驗中，我們對近年著名的摘要研究進行比較。首先，指針生成網路(Pointer Generator Network, PGN) (See *et al*., 2017)是相當知名的重寫式摘要法，但其結果甚至較抽取式的領導摘要法(LEAD)更差，這是許多重寫式摘要系統的缺點，即容易生成語意不通或不流暢的摘要，導致評估結果不佳。REFRESH (Narayan *et al*., 2018)提出一個使用加強式學習的抽取式摘要模型，其任務成效表現優於指針生成網路(PGN)以及領導摘要法(LEAD)。由於 BERT 的提出，許多自然語言處理的相關問題皆獲得大幅度的進步，在抽取式摘要的研究中，BERTSUM 是基於 BERT 而提出的經典方法，由實驗結果可知，他可較領導摘要、指針生成網路以及 REFRESH 有更佳的 ROUGE 分數。雖然 BERTSUM 已取得大幅度的進展，但他沒有考量了句子在文章中的位置資訊、摘要模型與摘要評估方式的關係並不密切，而且須要倚賴額外的三連詞過濾法以避免冗餘資訊的選取。為了解決這些問題，我們提出了一套新穎的基於 BERT 的強健性抽取式摘要法 EBSUM，實驗結果顯示，EBSUM 確實相較於領導摘要、指針生成網路、REFRESH 以及 BERTSUM 有更佳的 ROUGE 分數。

表 3. 各式經典的摘要模型於 CNN/DailyMail 資料集上的實驗結果。
[Table 3. Experimental Results of the Classic Models in CNN/DailyMail]

	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	40.32	17.56	36.58
PGN (See, Liu, & Manning, 2017)	39.53	17.28	37.98
REFRESH (Narayan, Cohen, & Lapata, 2018)	41.00	18.80	37.70
BERTSUM (Y. Liu, 2019)	43.25	20.24	39.63
EBSUM	**43.42**	**20.40**	**39.78**

表 4. 探究本研究所提出的各個改進元件之實驗結果。
[Table 4. Experimental Results of Improved Components in CNN/DailyMail]

	ROUGE-1	ROUGE-2	ROUGE-L
BERTSUM-Trigram Block	42.56	19.96	39.01
BERTSUM-Trigram Block+CLS	42.64	20.05	39.11
BERTSUM-Trigram Block+SEP	42.68	20.08	39.14
BERTSUM-Trigram Block+ALL	42.25	19.72	38.67
BERTSUM-Trigram Block+SEP+RL	42.71	20.14	39.18
EBSUM	**43.42**	**20.40**	**39.78**
EBSUM+Trigram Block	43.28	20.16	39.61

最後，我們逐一的探究本研究所提出的各個改進元件。首先，為了驗證句子位置向量對模型的影響，我們在不使用三連詞過濾法的 BERTSUM 中，表示為”BERTSUM-Trigram Block”，分別加入 CLS 向量、SEP 向量以及 ALL 向量，實驗結果如表 4 所示，分別表示為”BERTSUM-Trigram Block+CLS”、”BERTSUM-Trigram Block+SEP”與”BERTSUM-Trigram Block+ALL”。由實驗結果可知，將句子位置特徵加入所有字詞當中的表現最差（即 ALL 向量），這可能是因為將句子位置資訊加入所有字詞向量中，使得字詞向量表示法過於含糊，因此若僅將句子位置資訊加入[CLS]或[SEP]中，較不會影響字詞本身的向量表示法，且可以確實的將句子的位置資訊融入摘要的選取之中。接著，我們探究強化學習對摘要模型的影響，實驗結果如表 4 中”BERTSUM-Trigram Block+SEP+RL”所示。結果顯示，加入強化學習，確實可以有效地考量摘要的評估結果於模型之中，因此相較於”BERTSUM-Trigram Block+SEP”，可獲得一定的成效提升。最後，當我們於”BERTSUM-Trigram Block+SEP”的設定中再加入最大邊緣相關性向量，即成為本研究所提出之基於 BERT 的強健性抽取式摘要法(EBSUM)，由於融合了句子的位置資訊、強化學習以及最大邊緣相關性準則，因此 EBSUM 可以獲得最佳的摘要成效。當我們進一步地將 EBSUM 再與三連詞過濾法相結合時(即”EBSUM+Trigram Block”)，摘要的成績是下降的，這是因為最大邊緣相關性準則已自動地會將冗餘的資訊屏除，若再通過規則式的三連詞過濾法，可能會過度的將一些值得選取得句子摒棄，而造成任務成效降低的結果。

6. 結論 (Conclusions)

在本文中，我們提出了一套新穎的基於 BERT 之強健性抽取式摘要法 EBSUM，不僅考量了句子的位置資訊、利用強化學習增強摘要模型與評估標準的關聯性，更直接地將最大邊緣相關性概念融入摘要模型之中，因此在公開的評測資料集 CNN/DailyMail 中，EBSUM 可以獲得最佳的摘要任務成效。未來，我們將持續改進 EBSUM 的模型架構，使其可以更簡單、更有效，也希望驗證此一模型可以在不同語言的資料集中，展現摘要的成效。此外，我們也將把 EBSUM 應用於其他任務之中，諸如資訊檢索與機器翻譯等。

致謝 (Acknowledgements)

This work is supported by the Ministry of Science and Technology (MOST) in Taiwan under grant MOST 108-2636-E-011-005 (Young Scholar Fellowship Program), and by the Project J367B83100 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan.

參考文獻 (References)

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. In arXiv preprint arXiv:1607.06450

Billawala, Y., Mehdad, Y., Radev, D., Stent, A., & Thadani, K. (2018). Scalable and effective document summarization framework. In: Google Patents.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1-7), 107-117. doi: 10.1016/S0169-7552(98)00110-X

Carbonell, J. G., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceeding of SIGIR '98*, 335-336. doi: 10.1145/290941.291025

Cheng, J., & Lapata, M. (2016). Neural summarization by extracting sentences and words. In arXiv preprint arXiv:1603.07252.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In arXiv preprint arXiv:1810.04805

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, *22*(1), 457-479. doi: 10.1613/jair.1523

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Commun. Acm*, *57*(6), 74-81. doi: 10.1145/2602574

Gehrmann, S., Deng, Y., & Rush, A. M. (2018). Bottom-up abstractive summarization. In *Proceedings of EMNLP 2018*, 4098–4109. doi: 10.18653/v1/D18-1443. In arXiv preprint arXiv:1808.10792

Gu, Y., & Hu, Y. (2019). Extractive Summarization with Very Deep Pretrained Language Model. *IJAIA*, *10*(2), 27-32. doi: 10.5121/ijaia.2019.10203

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(2-3), 146-162. doi: 10.1080/00437956.1954.11659520

He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, *52*(7), 801-812. doi: 10.1016/j.im.2015.04.006

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M.,…Blunsom, P. (2015). Teaching machines to read and comprehend. In *Proceedings*

of the 28th International Conference on Neural Information Processing Systems, *1*, 1693-1701.

Hirao, T., Yoshida, Y., Nishino, M., Yasuda, N., & Nagata, M. (2013). Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1515-1520.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *CSUR*, *47*(4), 67. doi: 10.1145/2771588

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017*, 67-72. In arXiv preprint arXiv:1701.02810

Lebanoff, L., Song, K., Dernoncourt, F., Kim, D. S., Kim, S., Chang, W., … Liu, F. (2019). Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of ACL 2019*, 2175-2189. In arXiv preprint arXiv: 1906.00077

Leiva, L. A. (2018). Responsive snippets: adaptive skim-reading for mobile devices. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. doi: 10.1145/3236112.3236159

Li, P., Wang, Z., Lam, W., Ren, Z., & Bing, L. (2017). Salience estimation via variational auto-encoders for multi-document summarization. In *Proceedings of Thirty-First AAAI Conference on Artificial Intelligence*, 3497-3503.

Lin, C.-Y., & Hovy, E. (2002). From single to multi-document summarization. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, 457-464. doi: 10.3115/1073083.1073160

Liu, F., Flanigan, J., Thomson, S., Sadeh, N., & Smith, N. A. (2018). Toward abstractive summarization using semantic representations. In arXiv preprint arXiv: 1805.10399

Liu, L., Lu, Y., Yang, M., Qu, Q., Zhu, J., & Li, H. (2018). Generative adversarial network for abstractive text summarization. In *Proceedings of Thirty-second AAAI conference on artificial intelligence*, 8109-8110.

Liu, Y. (2019). Fine-tune BERT for Extractive Summarization. In arXiv preprint arXiv: 1903.10318

Marc, D. (1998). Improving summarization through rhetorical parsing tuning. In *Proceedings of Sixth Workshop on Very Large Corpora*, 206-215.

Maskey, S. R., & Hirschberg, J. (2003). Automatic summarization of broadcast news using structural features. In *Proceedings of Eighth European Conference on Speech Communication and Technology*. doi: 10.7916/D8348TR5

Maybury, M. T., & Merlino Jr, A. E. (2005). Automated segmentation, information extraction, summarization, and presentation of broadcast news. In: Google Patents.

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 404-411.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In arXiv preprint arXiv: 1301.3781

Nallapati, R., Zhai, F., & Zhou, B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 3075-3081.

Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. In arXiv preprint arXiv: 1602.06023. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 280-290. doi: 10.18653/v1/K16-1028

Narayan, S., Cohen, S. B., & Lapata, M. (2018). Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), 1747-1759. doi: 10.18653/v1/N18-1158. In arXiv preprint arXiv: 1802.08636

Nogueira, R., & Cho, K. (2019). Passage Re-ranking with BERT. In arXiv preprint arXiv: 1901.04085

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREc' 10*.

Parveen, D., Ramsl, H.-M., & Strube, M. (2015). Topical coherence for graph-based extractive summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1949-1954. doi: 10.18653/v1/D15-1226

Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. In arXiv preprint arXiv:1705.04304.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. doi: 10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In arXiv preprint arXiv:1802.05365

Polajnar, T., Rimell, L., & Clark, S. (2015). An exploration of discourse-based sentence spaces for compositional distributional semantics. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, 1-11. doi: 10.18653/v1/W15-2701

Qu, C., Yang, L., Qiu, M., Croft, W. B., Zhang, Y., & Iyyer, M. (2019). BERT with History Answer Embedding for Conversational Question Answering. In arXiv preprint arXiv:1905.05412

Radev, D. R., Jing, H., & Budzikowska, M. (2000). Centroid-based summarization of multiple documents: sentence extration, utility-based evalutation, and user studies. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization*, 21-30. doi: 10.3115/1117575.1117578

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Retrieved from URL https://s3-us-west-2. amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic biology*, *67*(5), 901-904. doi: 10.1093/sysbio/syy032

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In arXiv preprint arXiv:. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). doi: 10.18653/v1/P17-1099

Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: a social media analytics framework. *Social network analysis and mining*, *3*(4), 1277-1291. doi: 10.1007/s13278-012-0079-3

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. Cambridge, MA: MIT press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the Advances in neural information processing systems(NIPS 2017)*, 5998-6008.

Wan, X. (2008). An exploration of document impact on graph-based multi-document summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP' 08)*, 755-762. doi: 10.3115/1613715.1613811

Wan, X., & Yang, J. (2008). Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 299-306. doi: 10.1145/1390334.1390386

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, *8*(3-4), 229-256. doi: 10.1007/BF00992696

Wong, K.-F., Wu, M., & Li, W. (2008). Extractive summarization using supervised and semi-supervised learning. In *Proceedings of COLING 2008*, 985-992. doi: 10.3115/1599081.1599205

Wu, Z.-Y., Yen, L.-P., & Chen, K.-Y. (2019). Generating Pseudo-relevant Representations for Spoken Document Retrieval. In *Proceedings of ICASSP 2019*. doi: 10.1109/ICASSP.2019.8683832

Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., . . . Lin, J. (2019). End-to-end open-domain question answering with bertserini. In arXiv preprint arXiv: 1902.01718. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*.

Yang, W., Zhang, H., & Lin, J. (2019). Simple applications of bert for ad hoc document retrieval. In arXiv preprint arXiv: 1903.10972

Yin, W., & Pei, Y. (2015). Optimizing sentence modeling and selection for document summarization. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 1383-1389.

Zhang, H., Gong, Y., Yan, Y., Duan, N., Xu, J., Wang, J., . . . Zhou, M. (2019). Pretraining-Based Natural Language Generation for Text Summarization. In arXiv preprint arXiv:1902.09243

Zheng, H., & Lapata, M. (2019a). Sentence Centrality Revisited for Unsupervised Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Zheng, H., & Lapata, M. (2019b). Sentence Centrality Revisited for Unsupervised Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.