

# 2019 年臺灣國際科學展覽會 優勝作品專輯

作品編號 190005

參展科別 電腦科學與資訊工程

作品名稱 英文句子依閱讀程度進行簡化之研究

得獎獎項 大會獎：四等獎

候補作品

就讀學校 臺北市立第一女子高級中學

指導教師 黃芳蘭、陳信希

作者姓名 戴雅婕

關鍵詞 語言學習、句子改寫、深度學習

## 作者簡介



大家好，我是北一女中三年級的戴雅婕，自國小讀資優班起，我就對資訊的廣泛應用、主題的深入探討產生濃厚的興趣，尤其在高中的專題課程，經過緊密的接觸、繁複的實驗，不斷的分析中彷彿開啟了心中對資訊探索的黑盒子，一頭栽進主動學習的領域，享受探索過程中的各種考驗。

這次研究，我要由衷的感謝在此研究之旅中一路陪伴與支持我的家人、師長與同儕，期待自己未來亦能一秉初衷的將對此研究的熱忱，陸續發現應用生活之鑰，打開智慧生活之門。

# 摘要

英文句子簡化是一項單語言句子轉換的任務，其中一句複雜的句子會轉換為一句或多句的簡單句子。相較於過去研究學者著重於研究如何優化句子簡化的結果，如何將一句英文句子依閱讀程度簡化為不同簡單程度的簡化句是一項自然語言處理方面嶄新的研究領域。本研究首先訂定英文分級標準，整合歐洲(CEFR)與台灣(LTTC)母語非英語國家機構對英文的分級標準，將英文分為三種難易程度，並依此將 Wikipedia 及 Newsela 的簡化前-簡化後平行語料重新刪整為三種目標程度等級的平行語料庫。另一方面，運用已發展成熟的 Seq2seq 簡化模型，創造一個多解碼器模型，分別依據目標程度不同的訓練資料集訓練三種解碼器。在 BLEU、SARI 指標以及 Coverage 計算下，本研究結果相較於相關研究可展現出優異成果。

## ABSTRACT

Text simplification is a monolingual text-to-text transformation task where a complicated text is transformed into a simpler text. Compared to most recent work studying how to perform well in TS, our goal is to simplify sentences based on Reading Levels and output multiple simplifications of the same original text targeting different levels. We present the first attempt at classifying parallel corpus according to 3 reading levels. Our approach uses a sequence-to-sequence architecture where we make it possible to simplify an original sequence into sequences of different lower levels. We show that it outperforms state of-the-art TS approaches similar to us.

# 壹、前言

## 一、研究動機

在英文普及與網際網路廣為運用的學習環境中，隨著使用的對象、目的不同，針對相同主題，環境如網際網路等常充斥內容深淺不同的資訊提供廣泛的使用者搜索。然而當單一使用者希望快速在環境中汲取各種資訊的內容時，卻常可能因自身能力的限制而受到汲取資訊範圍的侷限，「面對如此龐大的資料量，我們是不是能透過某種科學方法，將資訊快速改寫、使我們能有效從中學習呢？」我想，同時今天日益茁壯發展的資訊科技正具備不須人工逐一處理的特性，因此我希望結合資訊科技與日常生活、一步一步探索資訊科學領域中對於英文句子簡化改寫的方法，以達成「學以致用」的學習真諦。

## 二、研究目的

探討如何運用資訊科技簡化英文句子，並運用現有的資訊科技探索將相同的英文句子有效地簡化為不同程度的簡化句。

## 三、研究問題

(一) 藉由比較簡化前與簡化後的英文文章，探討英文文章簡化與英文句子簡化的可行性並確認研究發展方向。

表 1 摘錄於維基百科關於籃球運動發明之歷史，左欄是簡化前的內容 (<https://en.wikipedia.org/wiki/Basketball>)，右欄是在 Simple English 人工簡化後的版本 (<https://simple.wikipedia.org/wiki/Basketball>)。左欄中藍色文字為簡化後省略的部分，兩欄中紅色文字則為改寫之前後內容。例如左欄中 [the International Young Men's Christian Association Training School \(YMCA\) \(today, Springfield College\)](#)，在右欄省略為 [Springfield](#)

College。而右欄中 professor(教授)簡化後則改寫為 teacher(老師), so the bottom of the basket was removed ,allowing the balls to be poked out with a long dowel each time.更改寫為 people made a hole at the bottom of the basket so the ball could go through more easily.。

由以上的例子得知，進行篇章簡化改寫包括：

- 1．決定哪些句子片段是重要的，要保留進行改寫。哪些句子片段相對不重要，可以省略。
- 2．詞彙層次的改寫：決定哪個詞彙要被改寫，要用哪個詞彙來改寫。例如：根據大學入學考試中心《高中英文參考詞彙表》之編輯方法原則，professor 屬於第四級，改寫後的 teacher 則屬於最簡單的第一級。
- 3．結構層次的改寫：以常用的文法結構，替代複雜的結構。

由以上分析，篇章改寫牽涉到概念簡化等議題，這一次研究集中於句子改寫的問題，其成果可以做為篇章簡化研究之基礎，而如何有效對句子進行簡化，並產生不同程度的簡化句子即為本研究之研究目標。

表 1

文本簡化前後範例

簡化前	簡化後
In early December 1891, Canadian Dr. James Naismith, a physical education professor and instructor at the International Young Men's Christian Association Training School (YMCA) (today, Springfield College)in Springfield, Massachusetts,	In early December 1891, James Naismith, a Canadian physical education teacher at Springfield College in Springfield, Massachusetts, ... people made a hole at the bottom of the basket so the ball could go through more easily.

<p>...</p> <p>so the bottom of the basket was removed ,allowing the balls to be poked out with a long dowel each time.</p>	
--	--

- (一) 蒐集國內外公布的英文分級標準、由詞彙層面及結構層面訂定較適合國內英文閱讀者需求的分級指標對語料進行分級處理。
- (二) 探討如何運用、改寫序列到序列模型架構將一句輸入句簡化為多句不同簡單等級的輸出句。
- (三) 探討不同模型架構、參數設定與訓練資料多寡對於模型訓練的效能表現。

## 貳、研究方法或過程

### 一、 研究設備及器材

#### (一) 硬體

- 1 · 筆記型電腦 (寫程式。CPU：Intel(R)Core(TM)i7-6500U CPU @ 2.50GHz 2.60GHz)
- 2 · 工作站 (執行程式。CPU：Intel Core i7-2600K CPU @ 3.40GHz 四核心)

#### (二) 軟體及工具

- 1 · Python (程式語言)

一種為大眾廣泛使用的物件導向式高階程式語言。強調程式的可讀性與簡潔語法，同時擁有可跨平台執行的特性。

- 2 · Pytorch (深度學習框架)

一種以 Python 為優先考量的深度學習框架，可較快速搭建出模型。

### 3 · OpenNMT (開源神經機器學習翻譯系統)

一項自 2016 年推出的神經機器翻譯和神經序列建模的開源計劃，可用於文本摘要產生、圖像到文本、或語音到文本等領域。

### 4 · Stanford CoreNLP (核心自然語言處理工具)

一個自然語言處理工具，包含廣泛的語法分析工具，且能同時應用多個分析工具於一篇文本。

### 5 · NLTK (核心自然語言處理工具)

以 Python 處理自然語言數據的平台，提供多個語料庫、詞彙資源及用於詞性標記、語法分析等的文本處理庫。

### 6 · BLEU (機器翻譯評分工具)

### 7 · SARI (句子簡化評分工具)

## 二、實驗流程

首先，我們進行相關文獻探討，比較過去關於英文句子簡化研究的簡化方法並確立本研究的研究方向；接著蒐集包含簡化前英文句子與簡化後英文句子的平行語料，並參考、整合國內外英文相關單位(如：財團法人語言訓練中心、歐洲語言學習、教學、評量共同參考架構)對於英文的分級依據，訂定囊括詞彙層面、句法層面等兩層面的分級標準，淘汰不符合標準的平行語料，建立一個符合本研究宗旨的分級平行語料數據集；下一步，我們運用於相關文獻中效能較佳的研究方法，嘗試調整參數設定、改寫模型等多種方法，達成本研究有效簡化一個英文句子為多個簡化句子的目標，最後進行研究結果的討論與結論。

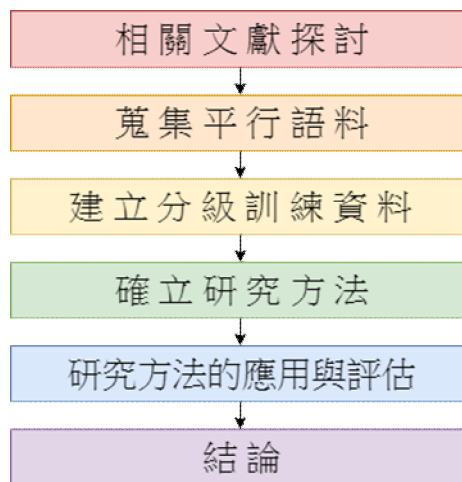


圖 1 研究流程圖

### 三、相關研究探討

英文句子簡化 (TS) 是一個近年來廣為研究的一項自然語言處理的領域，其目的在於：一個句子在經過字詞替換 (Rewording)、刪節 (Deletion)、重新排序 (Reorders)、合併 (Merges)、分句 (Split) (William Coster and David Kauchak, 2011) 等簡化過程後能轉換為另一句簡單句。由於將較難的英文句子轉換成較簡單的英文句子類似於英文句子翻譯，英文句子簡化任務亦被視為一種機器翻譯研究。

表 1 整理英文句子簡化相關文獻。起初，相關研究運用統計機器翻譯 (SMT) 處理英文句子簡化的詞彙、短語、句法等三種層面，然而根據基於短語的統計機器學習研究 (Specia, 2010) (Stajner et al., 2015) 以及基於句法的統計機器學習研究 (Xu et al., 2016)，運用統計機器學習會面臨字詞替換不當而導致文義前後不連貫的情況。Glavaš and Stajner (2015) 以及 Paetzold and Specia (2016) 創立另一個英文句子簡化的方法，運用無監督式方法的詞嵌入 (Word embedding) 詞彙簡化 (LS) 成功替換較難的字詞，但同時面臨因過多的字詞替換而產生語意模糊的困境。文本簡化神經網絡模型 (Neural Text Simplification Model) (Nisioi et al., 2017) 是一種能有效解決上述的問題的英文簡化方法，其特色在於 UNK 替換可以降低專有名詞對結果的影響，即使研究結果可能會因訓練資料多寡而產生影響，在網路發達、資訊共享的現今，訓練資料量並不會是一個無法克服的問題。



表 2

相關文獻方法比較

	方法	優點	缺點
PBSMT	基於短語的統計機器學習	可有效將難字替換為較簡單的字。	片語可能不當替換而導致文法有誤。
SBMT+SARI	基於句法的統計機器學習，目標設為 SARI 的呈現。	可避免上者句構、句義失當問題。	無法有效將難字替換為較簡單的字。
Light-LS	詞向量化及詞嵌入	不需有平行語料	可能替換為詞義相似之詞語導致意義分歧。
NMT	序列到序列神經機器學習。	1.輸入大量訓練資料可有效簡化且提高其準確度。 2.運用 UNK 替換，降低專有名詞對結果的影響。	訓練資料多寡可能影響結果。

由以上比較，NMT 是一種能有效進行英文句子簡化的模型，因此本研究將以此為基礎，進行一句英文句子簡化為多種程度英文句子的研究。

表 3

相關研究(Carolina Scarton、Lucia Specia, 2018)句子簡化配對範例

目標程度	美國二年級
標籤	

to-grade	<2> dusty handprints stood out against the rust of the fence near Sasabe.
operation	<elaboration> dusty handprints stood out against the rust of the fence near Sasabe.
to-grade-operation	<2- elaboration> dusty handprints stood out against the rust of the fence near Sasabe.
參考解答	dusty handprints could be seen on the fence near Sasabe.

表 3 摘錄自(Carolina Scarton and Lucia Specia, 2018) , Carolina Scarton and Lucia Specia 運用 OpenNMT (Klein et al., 2017) 開源神經機器學習系統，蒐集 Newsela 線上平台分享的文本資料，取其述說同一議題而適合美國二年級到十二年級學生閱讀的新聞句子作為平行語料（訓練資料集：440, 516 句及測試資料集：55, 064 句），接著統整四種句子簡化動作(Operation)：1 · Identical：一句輸入句對應一句與之相同的輸出句。2 · Elaboration：一句輸入句對應一句經簡化改寫的輸出句。3 · One-to-many：一句輸入句對應多句不同簡單程度的輸出句。4 · Many-to-one：多句不同程度的輸入句對應一句輸出句，最後在訓練資料集及測試資料集的輸入端句子分別加入 <to-grade>、<operation>、<to-grade-operation> 等標籤進行模型訓練。

然而， Newsela 公布的文本分級方法是以篇章為單位進行分級，因此在程度為美國四年級的文本中仍不免出現較低年級(如美國二、三年級)讀者也看得懂的句子，如此在訓練過程中仍將目標程度 grade 設為<4>即很可能模糊目標，最後導致句子目標程度僅成為一個宣稱的程度。

綜觀上述，本研究將以分級標準的訂定為首要目標，參考英文相關單位公布的分級參考依據，接著嘗試調整參數、改寫模型等方式，有效在將一句英文句子簡化為多種不同程度句子的目標下改善模型效能。

#### 四、 蒐集平行語料

##### (一) 維基百科及簡單英文維基百科 (W-SW)

維基百科是一個線上可供大眾共同參與、撰寫的百科全書，內容可有不同語言、不同英文程度的版本，本研究運用 W-SW 的特色，蒐集 Sergiu Nisioi、Sanja Stajner、Simone Paolo、Ponzetto 與 Liviu P. Dinu(2017)運用於序列到序列模型架構中的平行語料、以及 William Coaster 與 David Kauchak(2011)公布的平行語料，將其維基百科版本作為原始版 (Normal)，簡單英文維基百科版本作為簡化版 (Simple)。

表 4 摘錄於維基百科關於數字 2 的介紹句子，上欄是簡化前的內容 (<https://en.wikipedia.org/wiki/2>)，下欄是在 Simple English Wikipedia 人工簡化後的版本 ([https://simple.wikipedia.org/wiki/2\\_\(number\)](https://simple.wikipedia.org/wiki/2_(number)))。上欄中藍色文字為簡化後省略的部分，兩欄中紅色文字則為改寫之前後內容。例如上欄中 **The number two** 在下欄即省略為 **Two**。而上欄中 **properties (屬性、特性)** 經過簡化後改寫為 **meanings (涵義)**；**mathematics (數學)** 經過簡化後則改寫為 **math (數學)**。

表 4

文本簡化前後範例

原始版	The number two has many properties in mathematics .
簡化版	Two has many meanings in math .

##### (二) Newsela

一個提供適合美國二年級到十二年級學生閱讀的新聞的線上平台，多篇相同議題的新聞經由人工改寫可以有一到多種不同程度版本。有鑑於由其統整出的平行語料具備下列四種簡化處理的特色：Identical、Elaboration、One-to-many、Many-to-one (Carolina Scarton and Lucia Specia, 2018)，本研究運用其簡化處理方式為

One-to-many 的平行語料(共 218 個句子配對)做為測試資料。此外，相較於 W-SW，Newsela 文本句被較為生活化的特色。

## 五、 建立分級訓練資料

以詞彙層次與句法層次作為分級標準的設立基礎，整理國內外機構公開的英文詞彙與句法的分級標準，歸納出三級分級標準，其中：等級一（Level1）相當於國內小學程度；等級二（Level2）相當於國內中學生程度；等級三（Level3）相當於國內中學生以上程度。



### （一） 詞彙層次

#### 1. 蒐集分級詞彙

表 5 比較國內不同機構公布的分級詞彙集，左欄為大學入學考試中心公布的高中英文字彙<sup>[1][2]</sup>，右欄為財團法人語言訓練中心公布的全國英檢參考字表<sup>[3]</sup>。由於右欄的字表單字數較多且範圍較廣，故本研究採用右欄全國英檢參考字表<sup>[3]</sup>為本研究用於資料分級的字彙集，並以其中初級、中級、中高級作為本研究等級一、等級二、等級三。

表 5

#### 不同詞彙表比較

語料	大學入學考試中心高中英文字彙	全民英檢參考字表																
總數	6499 個單字	8163 個單字(相差 1664 個)																
分級	6 級(兩級為一組參考程度)	3 級(初、中、中高級)																
	 <p>7000單字分布表</p> <table border="1"> <thead> <tr> <th>參考程度</th> <th>單字數</th> </tr> </thead> <tbody> <tr> <td>中小學</td> <td>2156</td> </tr> <tr> <td>高中學測</td> <td>2179</td> </tr> <tr> <td>高中指考</td> <td>2164</td> </tr> </tbody> </table>	參考程度	單字數	中小學	2156	高中學測	2179	高中指考	2164	 <p>全民英檢單字分布</p> <table border="1"> <thead> <tr> <th>參考程度</th> <th>單字數</th> </tr> </thead> <tbody> <tr> <td>中小學</td> <td>2190</td> </tr> <tr> <td>高中</td> <td>2684</td> </tr> <tr> <td>大學</td> <td>3289</td> </tr> </tbody> </table>	參考程度	單字數	中小學	2190	高中	2684	大學	3289
參考程度	單字數																	
中小學	2156																	
高中學測	2179																	
高中指考	2164																	
參考程度	單字數																	
中小學	2190																	
高中	2684																	
大學	3289																	

## 2 · 訂定分級標準

由於在日常生活中，當句子出現一個超出閱讀者程度的單字且前後都沒有讀者認識的單字足以使讀者推敲其意義時，即使此一句子只含一個超出讀者程度的單字，讀者往往無法掌握整句話的涵義。故我們將分級標準訂為：除去專有名詞如人名、地名等的單字以及可能由上下文推敲其意義的名詞片語外，所有詞彙中出現程度最高者。

## 3 · 撰寫程式，將句子依詞彙分級標準分級

### (1) 整理可供查詢的字典

過濾文本中的雜訊（如：英文拼法、註解），接著按照字彙、詞類的分類標準將每一個單字本身及其詞性、等級輸入所包含的內容包括：詞性（pos）及等級（level）。最後整理出為一個可供查閱的字典，其中每項資料物件以詞彙為鍵、該詞彙之詞性與等級為其相對應的值：

```
{"a": {"pos": "art.", "level": 1}, "A.M.": {"pos": "adv.", "level": 1},
```

### (2) 進行句子分級

運用 StanfordCoreNLP 工具，排除詞性標記為連續兩個名詞以上名詞片語以及專有名詞辨識（NER）標記為人名、地名、組織、日期等名詞；再運用 NLTK 的 WordNetLemmatizer 將其他所有詞彙進行還原處理，最後將還原後的詞彙與查閱字典，以出現於字典的所有詞彙中最高級作為該句的等級。

表 6 為句子分級範例，首先在 StanfordCoreNLP 的詞性與 NER 標記下，Kaspar Hauser 因為有兩個連續的 NNP 且其 NER 為 PERSON，故可排除於分級依據之外，另一方面，Germany 因 NER 標記為 LOCATION，故亦排除於分級依據之外；接著，運用 NLTK 的 WordNetLemmatizer 將 was 還原為 be、lived 還原為 live，經過查閱(1) 整理出的字典，得知詞彙還原後的出現的最高等級為 Level1，故本範

例句的最終分級結果等級為 Level1。

表 6

句子分級範例

句子	Kaspar	Hauser	was	a	child	who	lived	in	Germany	.
詞性	<u>NNP</u>	<u>NNP</u>	VBD	DT	NN	WP	VBD	IN	NN	.
NER	PERSON	PERSON	O	O	O	O	O	O	LOCATION	.
原形	Omit	Omit	be	a	child	who	live	in	Omit	.
分級	Omit	Omit	1	1	1	1	1	1	Omit	.

## (二) 句法層次

### 1. 蒐集分級句法標準

在(一)我們蒐集國內英文相關單位(LTTC)公布的英文分級單字表作為詞彙層次的分級依據，然而在句法層次，該機構沒有公佈明確的句法分級依據，故本研究參考其與另一英文相關機構—歐洲共同語言參考標準(CEFR, Common European Framework of Reference for Languages: Learning, Teaching, Assessment)的測驗成績參考標準，依照相對應的等級，以 CEFR 公布的分級標準作為本研究句法層次的分級依據，並同時參考 Michael A. Covington、Congzhou He、Cati Brown, Lorina Naci 與 John Brown (2006)提出的英文句子分級標準，接著訂定句法層面的分級標準。

圖二摘錄於財團法人語言訓練中心的全民英檢各級測驗成績參照

(<https://www.gept.org.tw/Score/ScoreForm.asp>)，由表中可知，第一欄的初級可對應於 CEFR 的 A2-B1；第二欄的中級可對應於 CEFR 的 B1-B2；第三欄的中高級可對應於 CEFR 的 B2-C1。

全民英檢各級測驗成績參照  
GEPT Score Concordance

January 2017

聽力與閱讀測驗 Listening and Reading Tests

CEFR	初級 Elementary	中級 Intermediate	中高級 High-Intermediate	高級 Advanced
C1			210-240	180-240 150-179
B2		210-240	190-209 160-189	130-149
B1	210-240	190-209 160-189	140-159	
A2	190-209 160-189	140-159		

寫作測驗 Writing Tests

CEFR	初級 Elementary	中級 Intermediate	中高級 High-Intermediate	高級 Advanced
C1*			100	4-5 3
B2*		100	90-99 80-89	2
B1*	100	90-99 80-89	70-79	
A2*	90-99 70-89	70-79		

口說測驗 Speaking Tests

CEFR	初級 Elementary	中級 Intermediate	中高級 High-Intermediate	高級 Advanced
C1*			100	4-5 3
B2*		100	90 80	2
B1*	100	90 80	70	
A2*	90 80	70		

圖二

全民英檢各級測驗成績參照

由以上歸納，本研究將 CEFR 的句法分級標準中 A1、A2 作為本研究 Level1 的相對應；B1、B2 作為本研究 Level2 的相對應；C1、C2 作為本研究 Level3 的相對應。

(1) CEFR 分級標準

CEFR 是由歐洲委員會於 2001 年通過的一套適用於學習、教學、評估的建議標準，英文依程度高低可分為：C2、C1、B2、B1、A2、A1 等六級，其中 C2 程度最高，A1 則相反。在句法層面上，CEFR 公布了一套廣泛的自我程度評估表，另一方面，English Profile，在歐洲委員會的支持下由英國劍橋大學領導的相關機構，再依照 CEFR 公布的標準整理出較為明確的分級

依據。本研究運用上述明確分級依據，作為分級標準訂定的參考。

表7 摘錄自 English Profile 關於句法分級特徵的描述。第二列 A2 (Level1) 以簡單句為主；B1、B2 (Level2) 以虛主詞為首的句子、分詞構句為主；C1、C2 著重於不定詞的位置不同。

表 7

CEFR 句法分級依據

Level	特徵
A2	<p>Simple sentences</p> <p>Sentences with clauses joined by that</p> <p>Descriptive phrases introduced by a past participle</p> <p>Simple direct wh- questions</p> <p>Simple sentences using infinitives</p> <p>Other infinitives</p> <p>Some modals</p>
B1	<p>ing clauses</p> <p>Whose relative clauses</p> <p>Indirect questions</p> <p>Clauses with what as subject/object</p> <p>Verb+object+infinitive</p> <p>easy + infinitive</p> <p>Some complex auxiliaries ( eg. would rather, had better )</p> <p>Additional modal uses</p>
B2	<p>ing clause before the main clause</p> <p>It + verb + infinitive phrase</p> <p>Wh-clause as subject of main clause</p>



	Reported speech Lexically-specific verbs/adjectives + infinitive
C1	Lexically-specific verbs + object + infinitive Might for permission Fewer grammatical errors with agreement, countability or word formation
C2	Some new lexically-specific verbs + object+ infinitive Longer utterances with greater accuracy

( 2 ) Revised D-Level Scale

D-Level Scale 是由 Covington, et al. (2006)改良的一種英文句子分級標準，英文依程度高低可分為 Level0 至 Level7 等八種等級。

表 8 歸納 Covington, et al. (2006)的句法分級依據，經統整後大略可分為：Level0 至 Level2 以簡單句為主；Level3 至 Level4 以關係子句、that 同位語子句、含虛主詞的句子為主；Level5 至 Level7 以分詞構句為主。

表 8

句法分級依據(Covington, et al., 2006)

Level	特徵
Level0	Simple sentences (includes questions, sentences with auxiliaries)
Level1	Infinitive or -ing complement
Level2	Conjoined noun phrases
Level3	Relative (or appositional) clause modifying object of main verb Nominalization in object position

	Finite clause
Level4	Non-finite Complement with its own understood subject Comparative with object of comparison
Level5	Sentences joined by a subordinating conjunction Nonfinite clauses in adjunct (not complement) positions
Level6	Relative (or appositional) clause modifying subject of main verb Embedded clause serving as subject of main verb Nominalization serving as subject of main verb
Level7	More than one level of embedding in a single sentence

## 2 · 訂定分級標準

統整 1 ( 1 )、( 2 ) 的兩種分級依據，並結合自身學習歷程，表 9 訂定出最終的句法分級標準，其中僅以簡單的句法：簡單句、副句、合句構成的句子訂定為 Level1；以同位語子句為主構成的句子訂定為 Level2；以關係子句、分詞構句為主構成的句子以及包含上述兩種以上特徵的句子則訂定為 Level3。

表 9

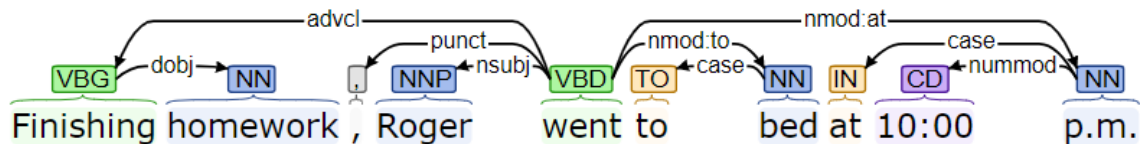
句法分級標準

分級	分級標準
Level1	簡單句、複句、合句
Level2	同位語子句
Level3	分詞構句、關係子句、含上述兩種特徵以上

## 3 · 撰寫程式，將句子依詞彙分級標準分級

運用 StanfordCoreNLP 的句法分析(Dependency trees)，觀察不同分級標準的句法分析

圖，並歸納其共通點，撰寫能偵測不同特徵的程式。



### (三) 建立分級數據集

分級數據集的建立與單字表標準的訂定方法相同，將（一）、（二）兩種層次下較高等級者作為該句的程度。

表 10 呈現數據集的句子分布。經由（一）詞彙層次、（二）句法層次的分級處理，W-SW 平行語料在各級的分布較為平均，Newsela 平行語料在各級的分布則較偏重於 Level3，由此可見以敘述句為主的 Newsela 平行語料因用語較生活化而導致句子可能因此非易懂的直述句。

表 10

依據詞彙與結構分級訓練資料比數

	W-SW 平行語料(句)	Newsela 平行語料(句)
Level1	53,169	13,334
Level2	58,413	29,774
Level3	51,954	78,395
Total	163,356	121,503

## 六、 確立研究方法

在相關研究探討中我們提到序列到序列模型架構是一目前在句子簡化領域中(Sergiu et al., 2017)能有較佳效能的深度學習模型架構。因此以下我們首先探討序列到序列模型原理及訓練過程，接著運用解碼器分級法改寫 OpenNMT 公布的開源程式碼模型展現出的效能，藉

以達成一句文句簡化成多種簡單程度文句。

### 1. 序列到序列模型(Sequence to sequence, Seq2seq)

序列到序列模型是一個建構於類神經網路之上的深度學習模型架構，首次被提出是被運用於機器翻譯領域上。直到 **Sergiu et al.(2017)**才近一步將其運用於句子簡化領域，並且效能還優於過去使用的統計機器學習方法。因此以下我們將探討。

由 **Sergiu et al.(2017)**運用的序列到序列模型，是將簡化前後的文本資料轉化為向量模式後，再作為雙層遞迴神經網路(RNN)的訓練資料訓練模型，其模型訓練過程架構如下。

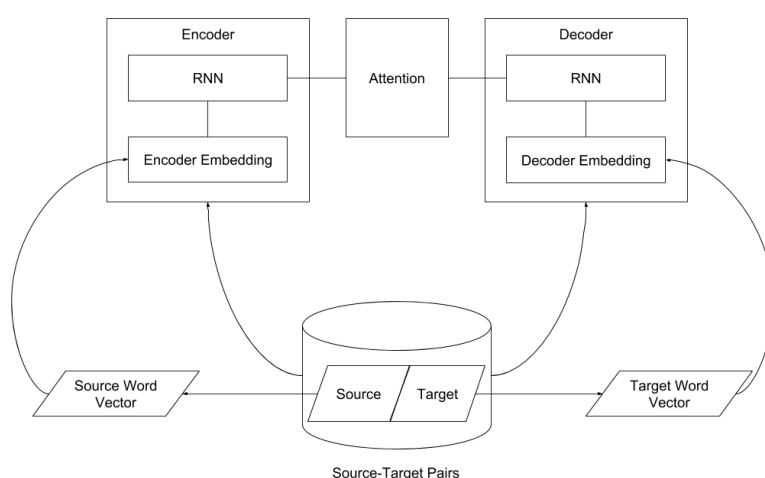


圖 3

### 序列到序列模型訓練架構

在圖 3 中，首先將訓練資料經過 **word2vec** 及 **word embedding** 向量化轉換後，將語句轉換為向量模式，再輸入編碼器-解碼器模型(Encoder-Decoder)並運用注意力機制(Attention)，最後經過模型的重複調校而訓練出數代模型。模型訓練結束後再運用測試資料評估出數代模型中表現最佳者，作為本模型訓練的最終參考結果。

#### (1) word2vec 及 word embedding

圖 4 我們以範例中簡化後的句子 **Two has many meanings in math.**舉例。我們可以發現：將詞語轉換為向量(**word2vec**)過程中，長度為 **n** 的語句會先以 **One-hot** 方式以一個 **n** 維向量呈現，再運用 **Skip-gram**、**CBOW** 等類神經網路方式轉換為一較緊密的

向量(word embedding)。其中 Skip-gram 是以目標詞語來預測目標詞語周圍的詞，CBOW 則相反。

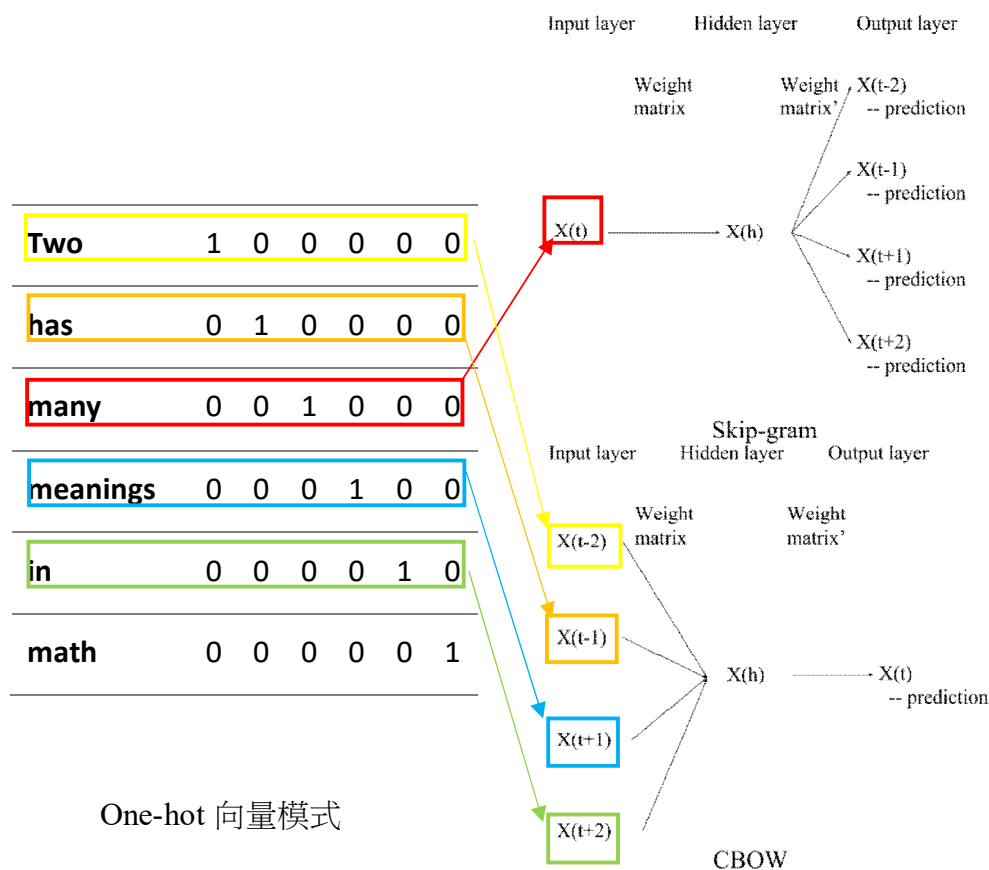


圖 4

## 詞語向量化

### (2) 編碼器解碼器模型及注意力機制

圖 5 中  $x$  代表輸入值， $o$  代表輸出值， $h$  代表隱藏狀態，而  $o$  與  $h$  相等。意即在 RNN 的訓練過程中，目前輸出向量是由前一個輸出向量與目前輸入向量一同計算而成的。由於此特色可使輸入與輸出語句之句長不相同並有效處理語句前後文連貫問題，故常用於自然語言處理領域。編碼器解碼器模型(Encoder-decoder)即運用 RNN 的此一特色，將輸入序列轉化為輸出序列。

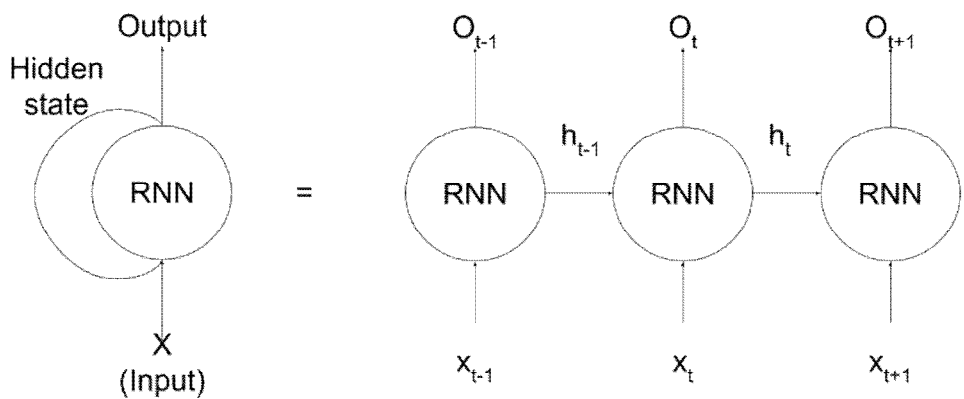
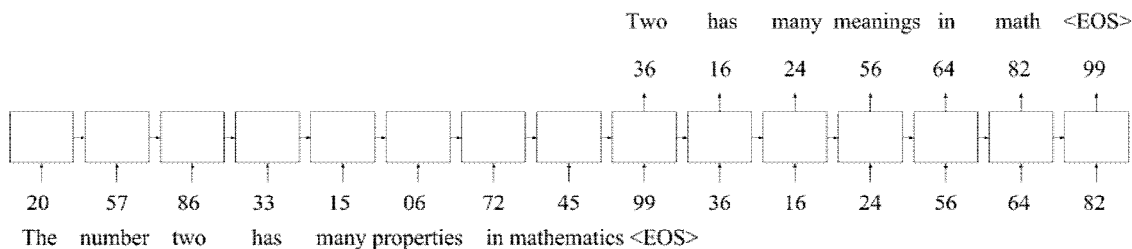


圖 5

### RNN 模型架構

圖 6 我們以簡化前後的語句進行舉例，訓練過程首先將簡化前語句轉換出的向量集合輸入 RNN 模型，壓縮為一個能表達整個語句意義的向量(Context vector)後，再透過另一 RNN 模型對其向量進行解壓縮，得出一個上下文可連貫向量集合，最後再運用 word embedding 找到訓練資料中簡化後的語句轉換出的向量集合進行相對應，進



而調教出最佳模型。

圖 6

### 編碼器解碼器訓練過程

#### 2. 解碼器分級法

改寫序列到序列模型架構，使訓練過程中編碼器與解碼器動態合作，訓練出能簡化一文句為多種簡單文句的模型。研究方法的應用與評估

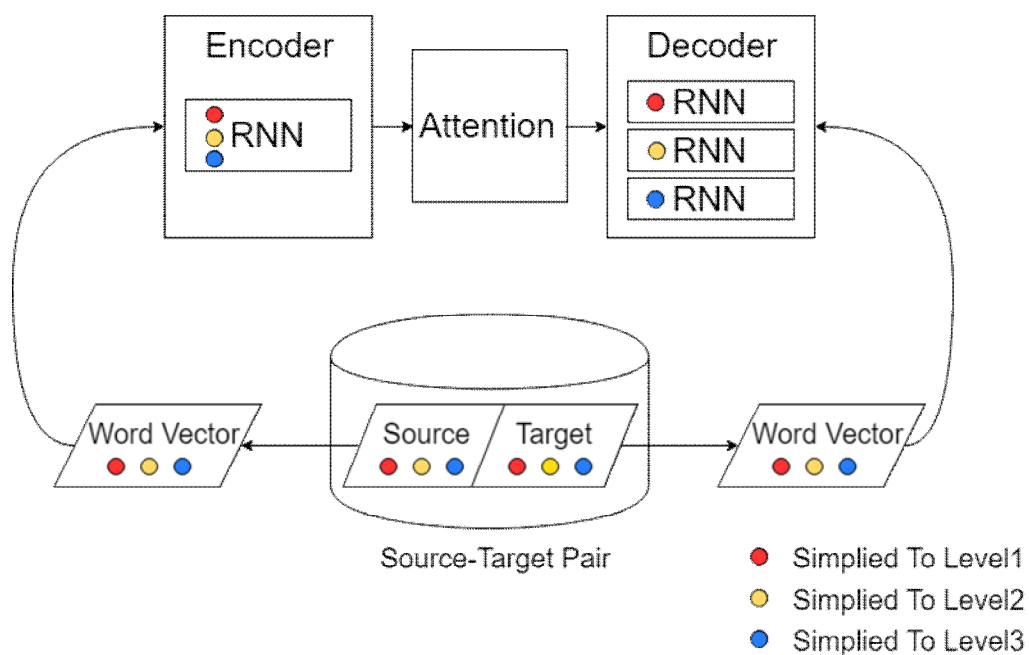


圖 7

多解碼器序列到序列生成架構

### 3.效能評估

#### (1) SARI

一種可呈現文句簡化程度的評分工具。將句子簡化結果輸出(Output)與原始版(Input)及相對原始版的簡化版參考解答(Reference)相互對應(圖 8)而得出 SARI 分數。

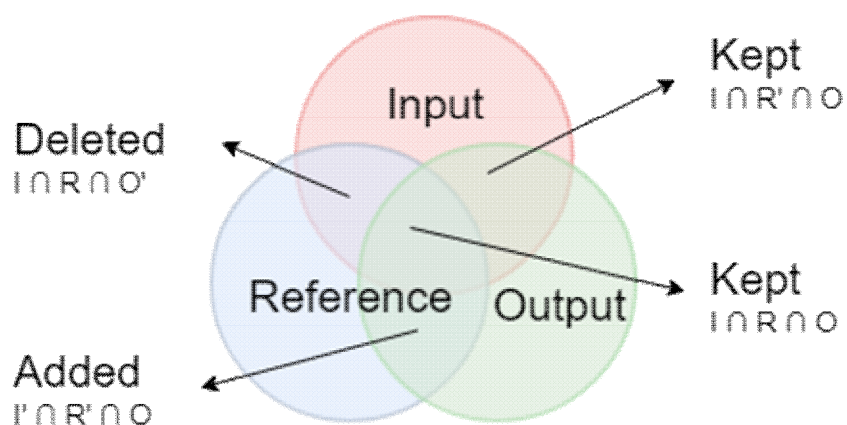


圖 8

## SARI 計算方式

### (2) BLEU

可有效呈現文句的通順程度的一種評分工具。分數由能展現機器翻譯結果(C)與參考解答彼此間的相似程度( $P_n$ )，與加權值(W)、長度懲罰因子(Brevity Penalty,

$$\text{BLEU} = \text{BP} * \exp\left(\sum_{n=1}^N W \log P_n\right)$$

BP)相乘得出。

#### i. 相似程度( $P_n$ )

$$P_n = \frac{\Sigma \text{Count}_{clip}(n - \text{gram} \text{ '})}{\Sigma \text{Count}(n - \text{gram})}$$

$$\text{Count}_{clip}(n - \text{gram}) = \min(\text{Count}(n - \text{gram}), \text{max\_Ref\_Count})$$

n-gram：長度為 n 的詞組。

n-gram'：第 k 個 n-gram 詞組。

$\Sigma \text{Count}(n - \text{gram})$ ：n-gram 在機器翻譯中出現的總次數。

$\text{Count}(n - \text{gram}')$ ：n-gram'同時出現在機器翻譯結果及參考解答中的次數。

max\_Ref\_Count：n-gram'在參考解答中出現的最大次數。

#### ii. 加權值(W)

W 為  $1/N$ ，N 為可取得的最大 n-gram。

#### iii. 長度懲罰因子(BP)

可避免長度較短、而只準確翻譯出部分句子的翻譯結果影響分數表現。因此若機器翻譯結果的句子長度(c)較參考解答的句子長度(r)小，BP 即會以較小值與其他參數相乘，達成抑制其句子影響分數之目的。

$$\text{BP} = \begin{cases} 1, & c > r \\ e^{(1-r/c)}, & c \leq r \end{cases}$$



### (3) 覆蓋率

一個可代表簡化成效的分數。計算方式將符合該程度的句子筆數與總測試資料筆數相除而得百分比。

$$\text{覆蓋率} = \frac{\text{符合該程度的句子筆數}}{\text{總測試資料筆數}} \times 100\%$$

下列以目標程度為 Level1 的 Output 文句計算為例：

Output 文句	分級	覆蓋率
Two has many meanings in math.	(1,1) → 1	$\frac{1}{3} \times 100\%$
The number two has many properties in mathematics.	(2,1) → 2	
Tallari had an outstanding season in Manchester, finishing as the EIHL 's second highest goal scorer, managing 55 goals and 38 assists in 59 games.	(2,3) → 3	

## 參、研究結果與討論

本研究運用訂定的分級標準篩選的 W-SW、Newsela 句子配對作為訓練資料集、Newsela 218 句可以一句原始句對應多句不同簡單程度的輸出句的句子配對作為測試資料集，比較本實驗 Multi-decoders 與 One Decoder、To-grade 等方法之效能。

1. One Decoder：運用序列到序列模型架構，訓練資料集為單一目標程度的句子配對。
2. To-grade<sup>[11]</sup>：運用序列到序列模型架構，訓練資料集為多種目標程度的句子配對，

並在每一組的原句前加上<目標程度>的標籤。

## 一、研究結果

表 1 1

不同模型在目標為 Level1 時的輸出表現

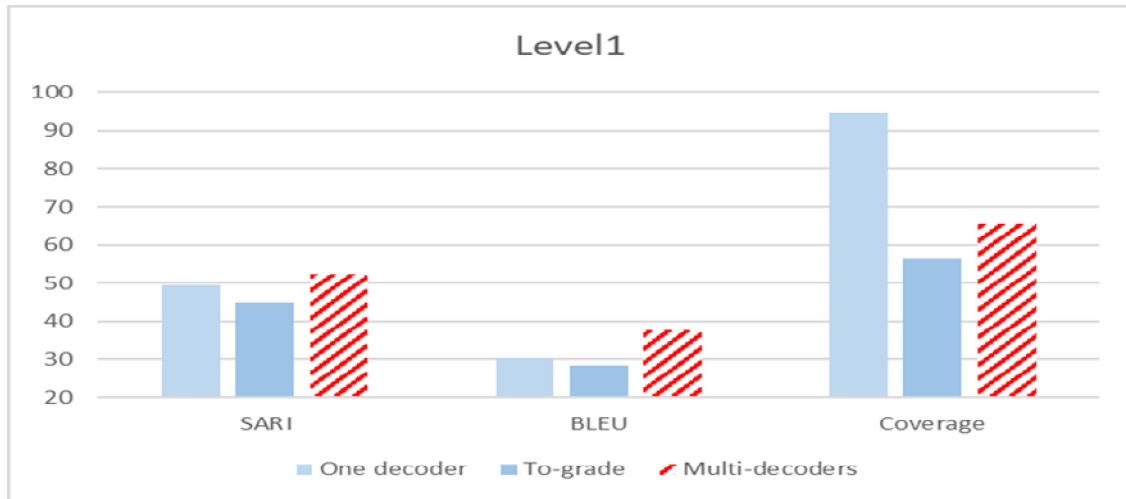


表 1 2

不同模型在目標為 Level2 時的輸出表現

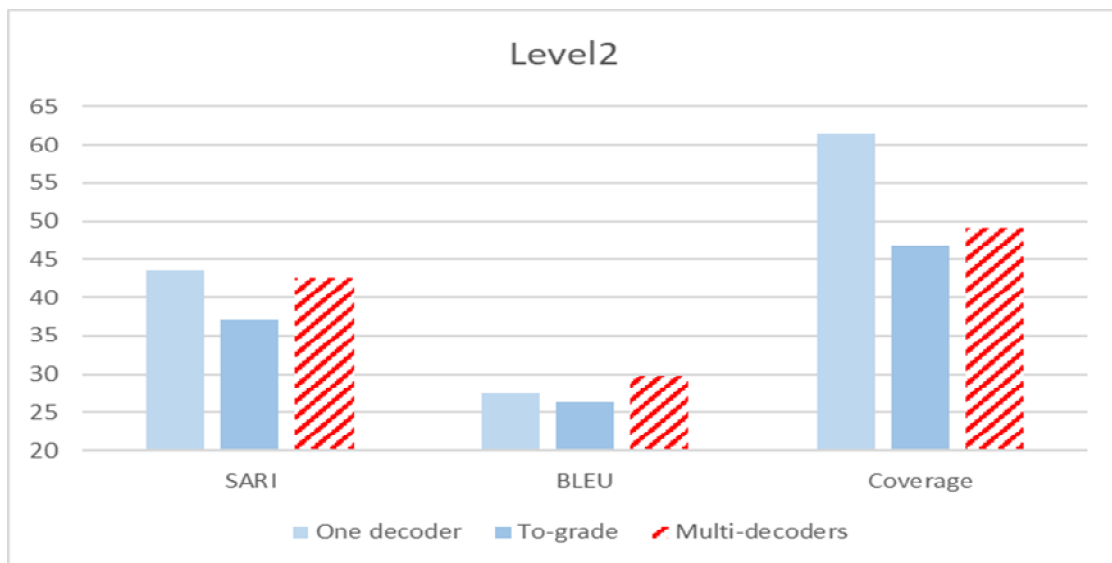
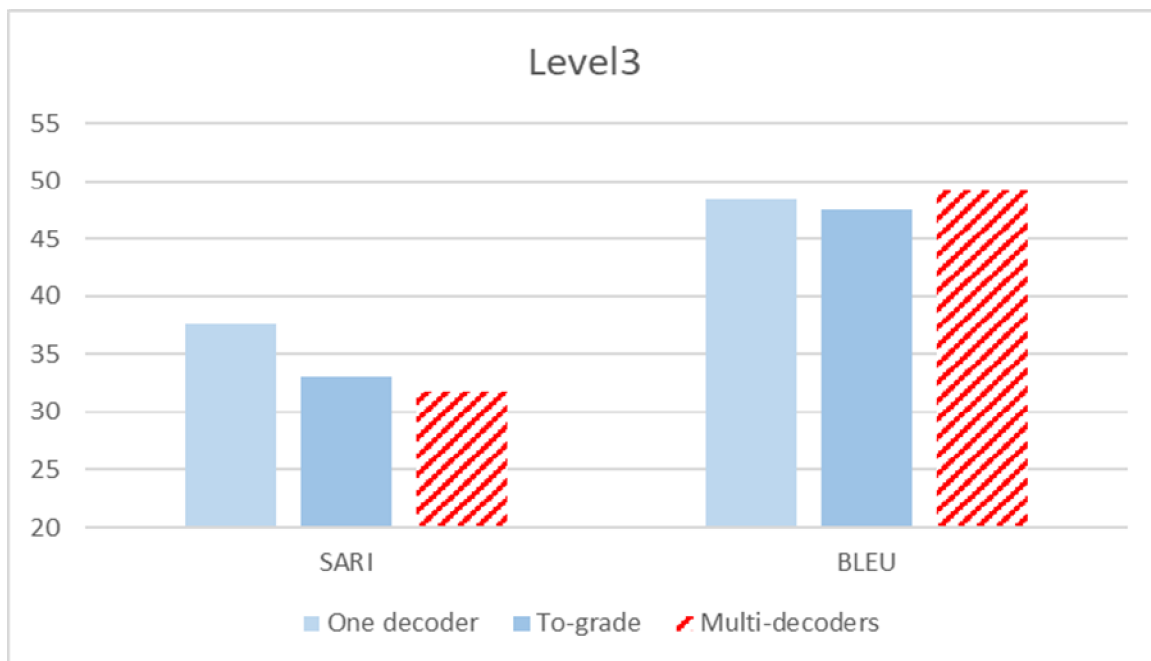


表 1 3

不同模型在目標為 Level3 時的輸出表現



## 二、 討論

### (一) SARI、BLEU 表現 Multi-decoders 佳

Multi-decoders 在共用同一個編碼器下，可學習較多元的語意表達方式，故在注重 Input、Reference 與 Output 彼此關係的機器評分工具表現上可得到較高的分數。

原句		Under the new policy , those two hours include using the Internet for entertainment , including Facebook , Twitter , TV and movies ; online homework is an exception .
Level1	Ref	<b>That would be</b> Facebook , Twitter , TV and movies .
	Output	<b>They</b> include Facebook , Twitter , TV and movies .
Level2	Ref	<b>Now</b> those two hours include using the Internet for entertainment
	Output	Under the new policy , those two hours include using the Internet for entertainment .

Level3	Ref	Under the new <b>guidelines</b> , those two hours include using the Internet for entertainment .
	Output	Under the new rules , those two hours include using the Internet for entertainment , including Facebook , Twitter , TV and movies .

## (二) Coverage 表現 One Decoder 大幅領先

One Decoder 運用單一目標程度句子配對作為訓練資料集，單字量較少，可生出較多符合範圍內的字詞，然而有鑑於識過的字詞除了較少、亦集中於單一程度，其輸出即可能出現謬誤而導致文義失真，推論 SARI、BLEU 等機器評分工具存在缺陷，無法有效評估文句的流暢度。

原句	But in its immediate wake , another problem formed, farmers without many resources were left with a lot of dead pigs .
One encoder	But in its <b>stages</b> , another problem formed .
Multi-decoders	But in its immediate wake , another problem formed .

## 肆、結論

### 一、研究結果優於相關文獻的成果

將英文句子依不同程度進行簡化在自然語言處理方面是較為嶄新的研究。

- 我們蒐集國內外英文分級的文獻、訂定分級標準→撰寫程式建立分級訓練資料→改寫序列到序列模型架構使模型能將一句語句簡化為三種不同等級的英文句子。
- 本研究改寫模型架構，運用 **Multi-decoders**，藉**共用編碼器**、學習多元的語意表達方式，成功在 SARI、BLEU 等的評分標準下展現較優異的簡化成果。

### 二、分級標準精準

- 相關研究[11]運用 **Newsela** 公布的文本資料作為訓練資料，與本研究間最大的差異在於「分級方法」。

- 本研究的分級方法是以「句子」為單位，而相關研究中則是採用「篇章」為單位進行分級，將文章中每一句話視為同一等級，可能模糊訓練資料目標程度。
- 本研究蒐集國內外各分級標準，為較精準的標準。

## 伍、未來展望

### 一、人工評分

由討論二得到：機器評分工具仍存在缺陷，在運算 Input、Reference、Output 之間的關係時，當一簡單字詞重複出現，即使文句的可讀性已喪失，評分工具仍可能將之評為高分，故希望加入人工評分，提升評分之可信度。

### 二、強化學習(Reinforcement Learning)

有鑑於目前模型訓練過程中，參數調整的同時無法得知翻譯的狀況，參考 Xingxing Zhang 與 Mirella Lapata(2017)，我們希望嘗試運用強化學習，將 SARI、BLEU、覆蓋率的表現納入序列到序列模行中 Loss 的運算，藉此在訓練過程中針對較佳的表現進行優化。

## 參考文獻

### 1.大學入學考試中心高中英文字彙參考表

[http://www.ceec.edu.tw/Research2/doc\\_980828/ce37/ce37.htm](http://www.ceec.edu.tw/Research2/doc_980828/ce37/ce37.htm)

### 2.國中英文課綱

[http://www.k12ea.gov.tw/97\\_sid17/%E8%8B%B1%E8%AA%9E970526%E5%AE%9A%E7%A8%BF%E5%96%AE%E5%86%8A.pdf](http://www.k12ea.gov.tw/97_sid17/%E8%8B%B1%E8%AA%9E970526%E5%AE%9A%E7%A8%BF%E5%96%AE%E5%86%8A.pdf)

### 3.財團法人語言訓練中心提供之全民英檢參考字表

<https://www.lttc.ntu.edu.tw/wordlist.htm>

### 4.國內英語能力比較檢測參考表

<http://intra.tpml.edu.tw/study/upload/downloads/table2.pdf>

### 5.CarolinaScarton and LuciaSpecia. 2018.” Learning Simplifications for Specific Target

- Audiences” In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pages 712-718
6. Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. ”Sequence to Sequence Learning with Neural Networks” In NIPS.
  7. Juri Ganikevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. “PPDB: The Paraphrase Database.” In Proceedings of NAACL-HLT 2013, pages 758-764
  8. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. ”BLEU: a Method for Automatic Evaluation of Machine Translation” In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311-318.
  9. Michael A. Covington, Congzhou He, Cati Brown, Lorina Naci, and John Brown. 2006. ”How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level scale” Research Report, AI Center, University of Georgia
  10. Sergiu Nisioi, Sanja Stajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. ”Exploring Neural Text Simplification Models” In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 85-91
  11. William Coaster and David Kauchak. 2011. “Simple English qWikipedia: A New Text Simplification Task.” In Proceedings of the 49th Annual Meeting of Association for Computational Linguistics, pages 665-669.
  12. Xingxing Zhang and Mirella Lapata. 2017. ”Sentence Simplification with Deep Reinforcement Learning” arXiv preprint arXiv:1703.10931 .

## 【評語】 190005

本作品針對英文句子簡化進行研究。運用 OpenNMT 之 Seq2seq 簡化模型架構，創造多解碼器模型，在一句英文句子以及目標程度的輸入下，輸出相對應程度的簡單英文句子，並與其他文獻比較簡化效能，資訊技術創新性佳，除了與既有文獻進行比較外，可多進行實用性實驗，以增加作品完整性。