國 立 中 正 大 學 語 言 學 研 究 所 碩 士 論 文

M.A. Thesis

Graduate Institute of Linguistics

National Chung Cheng University

人工智慧輔助華語教學：

以華語病句自動偵測、修正及建議為例

AI-assisted Chinese Language Teaching:

A Study of Ungrammatical Sentence Detection, Correction and Revision Suggestions

研究生：鍾芷軒

Graduate Student: Chih-Hsuan Chung

指導教授：吳俊雄 教授

Advisor: Prof. Jiun-Shiung Wu

中華民國 112 年 8 月

August 2023

致謝

回顧讀碩班的這三年，時光飛逝，每天都很忙碌也很充實，當我正在寫致謝的此時，代表我的碩士生涯已進入尾聲。對我而言，這份碩士論文充滿意義，我從沒想過自己會接觸程式語言，且能夠跨領域結合華語教學與自然語言處理，一步一步循序漸進完成華語病句自動偵測及修正系統——「華語吉 Bot 分」。

謝謝指導教授吳俊雄老師的用心教導，每當論文卡關時，吳老師總會適時指引思考方向，且大力提供相關協助，也謝謝鄂貞君老師和張瑜芸老師撥空前來擔任本次的口試委員，提供許多寶貴又實際的建議，讓我的碩士論文能更加完整。

因在課堂上接收到自然語言處理相關資訊，進而得知卓騰語言科技公司的實習計畫，我那時很猶豫是否要申請報名，我是華語教學的學生，對於自己的程式能力較沒信心，但我仍決定鼓起勇氣提出報名，並在 2022 年暑假時成為該公司的實習生。從實習過程中，我收穫良多，身旁的實習生都是臺灣頂尖國立大學的學生，處於如此積極正向的氛圍之下，有股動力督促著我必須更努力學習，就怕自己落後大家的進度。最後順利完成專題而取得實習證書，且延續專題內容而成為我的碩士論文，謝謝卓騰語言科技公司願意給我珍貴的實習機會。

這一路走來，受到很多人的照顧和幫助，所有的回憶我都謹記在心，並內心滿懷感恩。我的同學們書蘋、廷羽、品晴、紹華、梓靜，還有學妹們芯妍、珮瑄、孟婷，我們一起分享生活的快樂和讀書的苦悶，即使這段路途辛苦，但我們仍義無反顧追逐各自的夢想藍圖，由衷祝福日後一切順利，以及謝謝家人全力支持我，從來不要求我的課業成績，也不提及未來必須成功有所抱負，總跟我說：「放心去做你想做的事情，只希望你健康快樂就好！」

最後想謝謝自己，經歷無數疲憊的時刻，也時常懷疑自己的能力，卻從沒真正停下腳步，一直勇敢往前走著。我的內心有個信念，只要堅持著，不怕辛苦，總能走出一條屬於自己的道路。

摘要

現今華語教學領域強調教師的科技能力，華語也是自然語言處理領域相當關注的議題，因此，華語教學與自然語言處理跨領域的研究有其重要性。本論文以華語病句為研究重點，探討如何應用自然語言處理技術進行華語病句自動偵測及修正，並提供華語病句的錯誤說明。

自然語言處理領域大多把華語病句偵測視為二元分類，模型僅判斷句子是否為病句，無法完整解釋華語病句的錯誤為何，且大多把華語病句修正視為翻譯任務，也就是把病句翻譯成非病句，無法提供華語病句的錯誤說明。

本論文把華語病句偵測及修正視為分類任務，由模型進行一系列的病句相關判斷，逐步確認病句的錯誤分類，才能提供華語病句的建議說法和錯誤說明。有關模型架構，本論文比較三種模型，其中兩種為常見的機器學習模型，一為統計式簡單貝氏分類器(Naïve Bayesian Classifier)，二為深度學習神經網路長短程記憶模型(Long Short-Term Memory，簡稱 LSTM)，其中一種為基於語言學理論的模型，由卓騰語言科技公司(Droidtown Linguistic Tech. Co. Ltd.)開發的語言導向的關鍵詞介面(Wang et al. 2019，Linguistic Oriented Keyword Interface，簡稱 LOKI)。

本實驗分成訓練和測試兩個階段，並共有 640 個句子作為語料庫，從中分成訓練集、驗證集、測試集。病句語料來自美國某大學華語領航學程的課堂寫作，從中搜集 340 個病句；非病句語料來自國家教育研究院華語文語料庫與能力基準整合應用系統之華語中介語索引典系統，從中搜集 200 個非病句，故共有 540 個句子作為訓練階段的語料庫，依照自然語言處理的常用做法，分成 80%語料進行訓練，共有 432 個句子作為訓練集，而 20%語料進行驗證，共有 108 個句子作為驗證集。除此之外，本論文另外搜集 100 個句子作為測試集，主要用於測試階段，其中病句和非病句各 50 個，語料同樣來自國家教育研究院的華語中介語索引典系統。

在每個階段中，由模型經過三種分類，第一種分類為病句判斷，確認句子是否為病句；第二種分類為錯誤主類別判斷，錯誤主類別分成語法、語意、詞彙，由模型歸類病句屬於哪種錯誤主類別；第三種分類為錯誤次類別判斷，本論文採用 340 個病句，每個病句代表一種錯誤句型，故共有 340 個錯誤句型作為錯誤次類別，由模型歸類病句屬於哪種錯誤次類別。本論文逐步確認病句的錯誤分類，才能提供華語病句的建議說法和錯

誤說明。最後計算正確率(Accuracy)、精確率(Precision)、召回率(Recall)、F1 分數(F1-score)作為評估指標，在相同數量的少量語料下，比較三種模型的訓練表現及測試結果。

整體而言，LOKI 的各項評估指標皆優於簡單貝氏分類器和 LSTM，LOKI 的成績大致達到九成以上，以相同數量的少量語料而言，簡單貝氏分類器和 LSTM 無法如同 LOKI 有良好的表現，其原因為統計式和深度學習的模型需大量相同句型的句子，模型才能學習句型結構，但 LOKI 基於語言學的句法分析為運作原理，以詞組結構律(Phrase Structure Rules)為核心概念，其強大的特色在於以一個句子即可辨識相同句型的多個句子，實踐以少量語料即可辨識大量句子之目標。

華語病句的句型眾多，且結構複雜，LOKI 能辨識的句子為已建立的句型，僅需一個病句，即可學習其句型結構，本論文僅使用 340 個病句作為 LOKI 的語料庫，但 LOKI 能辨識的句子遠超過 340 個病句。最後，本論文討論三個華語病句相關議題，一為華語病句是否有規則，二為不同句型仍屬於同一種錯誤類別，三為錯誤類型的可能來源，以病句語料作為佐證，本論文初步推論華語病句可能有規則，且可能的原因為華語學習者的母語背景相同，此外，從本論文的病句語料中，發現不同的病句句型仍可歸類成相同的錯誤類別。

自然語言處理領域大多採用華語能力相關檢定的語料庫，代表研究所需的語料量大，但最關鍵的問題在於語料搜集不易，以及語料標註需花費大量的人力和時間，此外，目前電腦輔助語言學習領域有關華語病句的研究較少見。

本論文認為若能運用電腦科技自動偵測華語病句，且提供華語病句的建議說法和錯誤說明，以華語病句系統作為教學輔助工具，華語學習者使用該系統，不僅能即時得到華語病句修正的建議說法，也能從中初步了解病句的錯誤原因，這是一種自主學習的方式，將有助於實踐電腦輔助語言學習之目標，以提供華語教學相關領域作為參考。

關鍵字：華語教學、自然語言處理、電腦輔助語言學習、華語病句偵測及修正

Abstract

Nowadays, technological ability is very important to Chinese language teachers, and Chinese is also a topic of great concern to Natural Language Processing (NLP). Therefore, interdisciplinary research on Chinese Language Teaching and NLP is important and valuable. This study focuses on identifying and revising Chinese ungrammatical sentences, discusses how to apply NLP methods to automatically detect and correct Chinese ungrammatical sentences, and to provide Chinese language learners with error explanations.

NLP studies regard detection of Chinese ungrammatical sentences as a binary classification task, but such a task cannot explain why an ungrammatical sentence is ungrammatical. Moreover, correction of Chinese ungrammatical sentences is regarded as a translation task. In other words, an ungrammatical sentence is translating, by a model, into a grammatical one. But such a way cannot provide the reason why the sentences are ungrammatical.

In the thesis, the detection and correction of Chinese ungrammatical sentences are regarded as classification tasks. The models make a series of judgments about ungrammatical sentences. First, ungrammatical sentences are identified and distinguished from grammatical ones. Second, they are classified into one of three major error categories, which are syntactic error, semantic error, and vocabulary error, and then into one of the error subcategroies. Once we know their error subcategroies, we can provide revision suggestions and explanations for their errors.

This study compares three models. Two of the thtree are machine learning models -- Naïve Bayesian Classifier, and deep learning neural network Long Short-Term Memory (LSTM), and the other a linguistics-based model -- Linguistic Oriented Keyword Interface (LOKI) of Droidtown Linguistic Technology.

The experiment consists of two stages: training and testing. This thesis uses a total of 640 sentences, which is divided into the training set, the validation set and the test set. The corpus of ungrammatical sentences comes from the classroom writings of a university in the United States, where 340 ungrammatical sentences are collected; 200 grammatical sentences come from the Chinese Interlanguage Corpus of National Academy for Educational Research. The corpus of 540 sentences are used at the training stage. In addition, this thesis collects another 100 sentences for the testing stage. This corpus for the testing stage is composed of 50 ungrammatical sentences and 50 grammatical sentences, and is also collected from the Chinese Interlanguage Corpus of National Academy for Educational Research.

In each stage, the models perform three classification tasks as described above. The corpus of 540 sentences is uses at the training stage. 80% of it is the training set, and the other 20% is the validation set. The corpus of 100 sentences is used at the testing stage. Finally, the

performance of training is evaluated on the validation set and that of the trained model on the test set, with four indicators, i.e. Accuracy, Precision, Recall, and F1-score.

Overall, LOKI's performance in all four indicators is better than the Naïve Bayesian Classifier and LSTM. The reason for LOKI's good performance is that the statistical models and deep learning models need a large number of sentences so that the model can learn the sentence structures, but based on linguistics, LOKI relies on syntax to carry out the goal of recognizing multiple sentences based on a single input one. That is, LOKI can recognize a large number of sentences with a small amount of data.

Finally, this thesis discusses three more issues related to Chinese ungrammatical sentences: (1) whether there are rules that govern Chinese ungrammatical sentences, (2) different sentence patterns can still be of the same error type, and (3) possible sources of error types.

NLP studies mostly rely on Chinese proficiency test-related corpora, but the corpra used in such study have to be large, and it is labor-intensive and time-consuming to label large corpora. On the other hand, there is little research on Chinese ungrammatical sentences in the field of Computer-Assisted Language Learning (CALL). This study contributes in the sense that it relies on relatively small data and that it deals with automatically detecting and correcting Chinese ungrammatical sentences.

This study suggests that, if computer technology can be used to automatically detect Chinese ungrammatical sentences and to provide suggestions and explanations, using the AI-based Chinese ungrammatical sentence detection and revision system as a teaching assisting tool, Chinese language learners can not only immediately get suggestions for Chinese ungrammatical sentences, but also achieve self-learning.

Keywords: Chinese Language Teaching, Natural Language Processing, Computer-Assisted Language Learning, Chinese Ungrammatical Sentence Detection and Correction

目錄

圖目錄

表目錄

第一章 緒論

1.1 前言

有關華語教學與科技應用的研究日益廣泛（陳亮光 2008；鄭琇仁 2009；張于忻 2010；林翠雲 2012；詹衛東 2012；林翠雲 2013；陳姮良 2014；鄭琇仁 2014；詹衛東等人 2015；許德寶 2015；Lin et al. 2017；Tseng 2017；Hsin et al. 2017；Sung & Cheng 2017；李詩敏、林慶隆 2018；洪嘉馡等人 2018；連育仁 2018；Bai et al. 2019；曾妙芬等人 2019；Valdebenito & Chen 2019；Xu 2020；Tian 2020；洪嘉馡 2021；張莉萍 2022等），並興起一門學科稱為電腦輔助語言學習(Computer Assisted Language Learning，簡稱 CALL)，該領域的研究大致可分成華語師資培訓、華語教學課程、華語教材、華語教學課堂形式等類別。

為了提高華語教師的科技能力，目前有許多科技應用與華語師資培訓相關研究，例如陳亮光(2008)研究多媒體融入華語教學，以提升華語教師的提問技巧；鄭琇仁(2009)探討印尼地區的華語師資相關議題；林翠雲(2013)談論遠距教學與華語師資培訓等研究。

同時也有越來越多華語教師開始學習使用科技工具，並思考如何融入華語教學課堂之中，以增加學習效率，例如林翠雲(2012)探討 Web 2.0 數位工具於華語教學；陳姮良(2014)介紹如何透過教學平台和翻轉模式(Flipped models)來設計教學，讓華語學習者從中提高學習動機；洪嘉馡等人(2018)提出「中文階層式語法庫」作為輔助工具，協助學生完成自傳，並評估其華語寫作教學的成效等研究。

此外，華語教材不限於傳統紙本，例如張于忻(2010)基於模組化(Modulize)方式探討如何設計華語教材；連育仁(2018)研究如何以行動裝置整合語言教材作為數位化華語教材；洪嘉馡(2021)檢視如何建立數位平台輔助華語教學等研究。教學形式不只有實體課堂，也能運用視訊軟體進行線上課程，例如 Sung & Cheng (2017)談論華語教師和華語學習者對於視訊會議的線上教學之反饋等研究。華語教師的科技能力日漸重要，數位化多媒體已成為華語教學的發展趨勢。

除此之外，華語也是自然語言處理(Natural Language Processing，簡稱 NLP)領域相

當關注的議題(Yeh et al. 2016 ; Lee et al. 2017 ; He et al. 2018 ; Ren et al. 2018 ; Xiang 2018 ; Zhang & Wang 2019；Li et al. 2019；Zhao & Wang 2020；Wang et al. 2020；Cheng & Duan 2020；Lee et al. 2021；Ho et al. 2021；Chang et al. 2021；Chen & Zhang 2022；Kuang et al. 2022 等)，例如 He et al. (2018)採用卷積神經網路(Convolutional Neural Network，簡稱 CNN)模型辨識華語事件的真實性；Ho et al. (2021)採用 BERT (Bidirectional Encoder Representations from Transformers)模型及七種語言特徵來偵測臺灣華語的線上廣告業配文；Chang et al. (2021)運用統計方式，以銜接(cohesion)和連貫性(coherence)為研究重點，觀察這兩個語言特徵是否能辨識高影響力的社群媒體文章等研究。自然語言處理於華語教學的應用相關研究越來越多，採用各種模型來完成華語相關任務。

現今華語教學領域強調教師的科技能力，自然語言處理領域也相當關注華語相關議題，由此可見，華語教學與自然語言處理跨領域的研究有其重要性。本論文以華語病句為研究重點，探討如何應用自然語言處理技術進行華語病句自動偵測及修正之任務，並提供華語病句的錯誤說明。

本章架構如下：1.2 為研究目的及問題，介紹自然語言處理領域和電腦輔助語言學習領域相關研究，並提出本論文的研究問題；1.3 為研究架構，依序說明本論文的各章節內容。

1.2 研究目的及問題

在自然語言處理與華語病句相關研究中，可分為病句偵測和病句修正兩種，大多使用華語能力相關檢定的語料庫作為語料來源，並採用統計式或深度學習的模型，其研究所需語料量大。

統計式或深度學習的模型需大量相同句型的句子，模型才能學習句型結構，但最關鍵的難題在於語料搜集不易，以及語料標註需花費大量人力和時間，此外，自然語言處理領域大多把病句偵測視為二元分類，模型僅判斷句子是否為病句，無法完整解釋病句的錯誤為何，且大多把病句修正視為翻譯任務，也就是把病句翻譯成非病句，以序列對序列(sequence to sequence，簡稱 Seq2Seq)為主要的處理方式，雖然這種方式能達到病句修正的效果，且能直接更正其病句錯誤，但無法提供華語病句的錯誤說明。

以電腦輔助語言學習而言，有關運用各種科技工具於教學及輔助學習的研究日益廣

泛，研究大多以華語師資和教學課程為主軸，例如規劃科技應用的華語師資培訓課程(陳亮光 2008；鄭琇仁 2009；林翠雲 2012；鄭琇仁 2014；林翠雲 2013；Hsin et al. 2017；Tseng 2017；Lin et al. 2017 等)，以及採用多種科技工具來輔助華語教學課堂(陳姮良 2014；Sung & Cheng 2017；洪嘉馡等人 2018；Bai et al. 2019；Valdebenito & Chen 2019；曾妙芬等人 2019；Xu 2020；Tian 2020 等)，相較之下，有關華語病句的研究較少見。

本論文認為如果能運用電腦科技自動偵測華語病句，且提供華語病句的建議說法和錯誤說明，以華語病句自動偵測及修正系統作為教學輔助工具，華語學習者使用該系統，能即時得到華語病句修正的建議說法，並從中初步了解病句的錯誤原因，以提供華語學習者作為參考，這是一種自主學習的方式，將有助於實踐電腦輔助語言學習之目標。

本論文以華語病句為研究重點，語料搜集不易，故本實驗的樣本數量少。本論文選用三種模型，基於語言學理論的語言導向的關鍵詞介面(Linguistic Oriented Keyword Interface，簡稱 LOKI)、統計式簡單貝氏分類器(Naïve Bayesian Classifier)、深度學習神經網路長短程記憶模型(Long Short-Term Memory，簡稱 LSTM)，針對相同數量的少量語料進行參照比對，計算正確率(Accuracy)、精確率(Precision)、召回率(Recall)、F1 分數(F1-score)作為評估指標，比較此三種模型的表現差異，並從中應用自然語言處理技術，提供華語病句的建議說法和錯誤說明。

自然語言處理領域大多把華語病句修正視為翻譯任務，也就是把華語病句翻譯成非病句，但本論文則視為分類任務，逐步確認病句的錯誤分類，完成華語病句修正之任務。本實驗分成訓練和測試兩個階段，並在每個階段中，再細分三種分類，第一種分類為病句判斷，確認句子是否為病句；第二種分類為錯誤主類別判斷，確認病句屬於哪種錯誤主類別，並以語法、語意、詞彙此三種作為錯誤主類別；第三種分類為錯誤次類別判斷，確認病句屬於哪種錯誤次類別，每個病句代表一種錯誤句型，故共有 340 個錯誤句型作為錯誤次類別。

簡而言之，自然語言處理領域大多把病句偵測視為二元分類，僅判斷句子是否為病句，並沒有進行更詳細的錯誤分類，此方式無法完整解釋病句的錯誤為何，且大多把病句修正視為翻譯任務，以序列對序列為主要的處理方式，無法提供病句的錯誤說明，此外，該領域的研究大多以華語能力相關檢定的語料庫作為語料來源，其研究需大量語料，但最關鍵的難點在於語料搜集不易，以及語料標註需花費大量人力和時間；電腦輔助語言學習領域的研究大多以華語師資和教學課程為主軸，華語病句相關研究較少見。因此，

華語教學與自然語言處理跨領域的研究有其重要性，本論文運用自然語言處理技術達到華語病句自動偵測及修正之目標。

目前越來越多的研究證實，善用科技工具確實能提高語言學習的效率，故針對華語病句相關研究有其必要性，本論文的研究問題如下：

1. 比較基於語言學理論的 LOKI、統計式簡單貝氏分類器、深度學習神經網路 LSTM 此三種模型在華語病句自動偵測及修正的表現差異。
2. 在相同數量的少量語料下，LOKI、簡單貝氏分類器、LSTM 此三種模型的表現如何。
3. 針對少量語料的狀況，如何應用 LOKI 來開發華語病句自動偵測、修正及建議之人工智慧系統。

1.3 研究架構

關於本論文的架構，依序是第二章為文獻回顧，第三章為語料說明及實驗方法，第四章為研究結果及問題討論，第五章為結論與未來建議。

第二章為文獻回顧，將回顧有關自然語言處理技術於華語病句的應用及電腦輔助語言學習兩大類別的文獻。在自然語言處理領域中，有關華語病句的議題可分為病句偵測和病句修正。華語病句偵測大多為二元分類，無法完整解釋病句的錯誤為何，此外，通常把華語病句修正視為翻譯任務，無法提供病句的錯誤說明。該領域的研究大多使用華語能力相關檢定的語料庫，例如臺灣華語文能力測驗語料庫(Test of Chinese as a Foreign Language，簡稱 TOCFL)、中國大陸漢語水平考試語料庫(Hanyu Shuiping Kaoshi，簡稱 HSK)等，其研究所需的語料量大。以電腦輔助語言學習領域而言，規劃科技應用的華語師資培訓課程，以提高華語教師的科技能力，並思考如何運用多種科技工具於華語教學課堂之中，華語教材的呈現方式越來越多元，以語料庫作為教學素材的來源，建立語料庫作為教學的輔助工具，但有關華語病句的研究較少見。

第三章為語料說明及實驗方法，本論文的語料來自美國某大學華語領航學程的課堂寫作和國家教育研究院華語文語料庫與能力基準整合應用系統之華語中介語索引典系統，並介紹 LOKI、簡單貝氏分類器、LSTM 此三種模型的運作原理。本論文把華語病句修正視為分類任務，實驗分成兩個階段，依序為訓練階段和測試階段，訓練階段分成

訓練模型和驗證模型，完成訓練和驗證的實驗步驟，接續進入測試階段。在每個階段中，模型均進行三種分類，依序為病句判斷、錯誤主類別判斷、錯誤次類別判斷，並計算正確率、精確率、召回率、F1 分數作為評估指標，在相同數量的少量語料下，比較此三種模型的訓練表現及測試結果。

第四章為研究結果及問題討論，呈現 LOKI、簡單貝氏分類器、LSTM 此三種模型的訓練表現及測試結果，以及說明 LOKI 模型誤判的可能因素，並統整常見的病句錯誤，最後討論三個華語病句相關議題，一為華語病句是否有規則，二為不同句型仍屬於同一種錯誤類別，三為錯誤類型的可能來源，以病句語料作為佐證，本論文初步推論華語病句可能有規則，且可能的原因為華語學習者的母語背景相同，此外，從本論文的病句語料中，發現不同的病句句型仍可歸類成相同的錯誤類別。

第五章為結論與未來建議，總結本論文的內容，並提出未來的研究方向，以提供華語教學相關領域作為參考。

第二章 文獻回顧

2.1 概述

隨著科技日新月異的發展，現今人們藉由科技輔助提高各方面的效率，此概念同時發展於語言教學層面。從科技應用是否能結合語言教學已變成如何應用科技於語言教學之中(曾妙芬等人，2019)，科技與課堂的結合已是重要議題。

應用各種科技工具於教學或電腦輔助學習相關研究日益廣泛，例如觀察機器翻譯修正華語寫作、探討多種科技工具融入課堂之使用與成效、以訪談或問卷調查學生對於科技輔助學習的反饋等，目前有越來越多的研究證實，科技工具確實有助於提高語言學習的效率，此外，科技應用也被納入師資培訓的內容，因此，相關研究探討就職前的教師之科技實習計畫、師資的科技教育之現況和發展等。

本章關注兩大領域的研究，分別為自然語言處理技術於華語病句的應用及電腦輔助語言學習。在自然語言處理領域中，有關華語病句的議題可分為病句偵測和病句修正。華語病句偵測大多為二元分類，無法完整解釋病句的錯誤為何，此外，通常把華語病句修正視為翻譯任務，無法提供病句的錯誤說明；電腦輔助語言學習大多以華語師資和教學課程為主軸，有關華語病句的研究較少見。

本章架構如下：2.2 為自然語言處理技術於華語病句的應用，探討如何運用自然語言處理的方式於華語病句，可再細分二個面向，分別是華語病句偵測和華語病句修正；2.3 為電腦輔助語言學習，關注的重點在於電腦科技如何結合華語教學課堂，以達到輔助學習的成效；2.4 為結語，總結本章內容。

2.2 自然語言處理技術於華語病句的應用

語法錯誤修正(Grammatical Error Correction，簡稱 GEC)是當前自然語言處理領域相當關注的議題之一，英語的語法錯誤修正模型很多，相較之下，華語的模型較少，可能的原因在於華語的語料搜集不易、華語被視為困難的外語之一、語法規則複雜、用法多元等因素，華語學習者需具備相對較高程度，才足以有能力完成寫作，進而才能取得華

語的語料。

運用自然語言處理技術進行華語病句自動偵測及修正的研究越來越盛行，相關研究採用各種不同的神經網路模型來完成華語病句任務(Yeh et al. 2016；Lee et al. 2017；Ren et al. 2018；Xiang 2018；Zhang & Wang 2019；Li et al. 2019；Zhao & Wang 2020；Wang et al. 2020；Cheng & Duan 2020；Lee et al. 2021；Chen & Zhang 2022；Kuang et al. 2022 等)，而語料來源大致可分為繁體中文的臺灣華語文能力測驗 TOCFL 語料庫和簡體中文的中國大陸漢語水平考試 HSK 語料庫，並以病句偵測和病句修正為主要的研究方向。

在本節中，2.2.1 為華語病句偵測，大多把華語病句偵測視為二元分類，也就是由模型判斷句子是否為病句；2.2.2 為華語病句修正，大多把華語病句修正當作翻譯任務，其概念為把病句翻譯成非病句；2.2.3 為批判性回顧，提出本論文對於自然語言處理與華語病句相關研究的批判性回顧。

2.2.1 華語病句偵測

此類研究大多把語法錯誤視為二元分類，由模型判斷句子是否有語法錯誤，若句子有至少一個錯誤，即視為病句，以下為相關研究。

Yeh et al. (2016)採用 TOCFL 語料庫和 HSK 語料庫，並以臺灣中央研究院中文詞知識庫小組開發的 CKIP Tagger (Chinese Knowledge and Information Processing)進行斷詞，再把句子轉為詞向量，他們使用兩種深度學習神經網路模型，分別為循環神經網路(Recurrent Neural Network，簡稱 RNN) 和 LSTM，以 RNN-LSTM 作為模型架構，並於 TOCFL 語料庫和 HSK 語料庫分別進行病句偵測，以 TOCFL 語料庫而言，模型的表現為正確率 0.5218、精確率 0.5202、召回率 0.9726、F1 分數 0.6779；以 HSK 語料庫而言，模型的表現為正確率 0.5042、精確率 0.4964、召回率 0.9755、F1 分數 0.658。

Lee et al. (2017)採用 TOCFL 語料庫，該語料庫來自 46 種不同母語的華語學習者，總共 2837 篇作文，手動標記語法錯誤，結果取得 25277 個病句，68982 個非病句，並採用 CNN 和 LSTM 作為模型架構，從中比較 CNN、LSTM、CNN-LSTM 此三種模型的表現，最後 CNN-LSTM 的模型表現為正確率 0.6905、精確率 0.4439、召回率 0.5057、F1 分數 0.461。

Xiang (2018)採用 HSK 語料庫，並基於條件隨機場 (Conditional Random Field，簡

稱 CRF)作為模型架構，由模型進行三種分類，第一種分類為判斷句子是否有誤，第二種分類為辨識句子的特定錯誤類型，第三種分類為辨識錯誤位置，此類的模型是處理序列標註問題(sequence labeling problems)，可用於序列標註的語料，該研究的語法錯誤標籤設定為冗詞(redundant，代號 R)、缺失詞(missing，代號 M)、選詞錯誤(selection，代號 S)、詞序錯誤(word order，代號 W)、正確(correct，代號 O)，並分成一至六個字元處理，從中觀察模型的表現。關於模型分成六個字元的表現結果，第一種分類的成績為正確率 0.4559、精確率 0.6394、召回率 0.3081、F1 分數 0.4158；第二種分類的成績為正確率 0.3846、精確率 0.4003、召回率 0.1429、F1 分數 0.2107；第三種分類的成績為正確率 0.2652、精確率 0.0825、召回率 0.0227、F1 分數 0.0356。從結果顯示，第一種分類為判斷句子是否為有誤，屬於二元分類，而第二種分類為辨識錯誤類型，以及第三種分類為辨識錯誤位置，皆屬於多元分類，模型判斷會因分類難度增加而更容易形成誤判的狀況。

Zhang & Wang (2019)提到華語單詞使用錯誤(Chinese Word Usage Errors，簡稱 WUEs)常發生於華語學習者的寫作文章中，該研究的目標為自動偵測華語單詞錯誤，預測 WUEs 的位置及預測 POS 詞性標註，他們採用 HSK 語料庫，以並行雙向長短程記憶模型(Bidirectional Long Short Term Memory，簡稱 Bi-LSTM)作為模型架構，並結合多功能任務學習(multi-task learning)的方式進行訓練，只限於偵測單詞的使用錯誤，模型的正確率為 0.5338。

Cheng & Duan (2020)研究語法錯誤偵測，不區分中小學華語母語者和華語學習者的病句。針對華語作為外語(Chinese as a Foreign Language，簡稱 CFL)的華語學習者所寫的作文篇章，且運用自然語言處理技術偵測語法錯誤，使用三種語料庫，一為 HSK 語料庫，語料來自母語非華語的學習者參加漢語水平考試的寫作文章；二為 Lang-8 語料庫，這是一個語言學習網站，母語人士可自行選擇批改學習者的文章；三為課堂語料(School Dataset)，語料來自魯東大學(Ludong University)附設中小學，其作業、作文、週記、日記、試卷等。該研究把語法錯誤視為二元分類，以 BERT 作為模型架構，由模型判斷一個句子是否有語法錯誤，並從中比較三種 BERT 模型的表現，分別為 BERT-base、RoBERTa、 RoBERTa-wwm。該研究的錯誤類型分成四種，分別是冗詞(redundant words，代號 R)、缺失詞(missing words，代號 M)、選詞錯誤(word selection errors，代號 S)、詞序錯誤(word ordering errors，代號 W)，並計算 FP 值(False Positive Rate，簡稱 FPR)、精

確率、召回率、F1 分數作為評估指標，模型表現的成績分為驗證結果和測試結果，以 BERT-base 的成績而言，驗證結果為 FP 值 0.134、精確率 0.833、召回率 0.667、F1 分數 0.741；測試結果為 FP 值 0.052、精確率 0.98、召回率 0.685、F1 分數 0.807。

Lee et al. (2021)提到華語學習者可能出現漏詞(missing words)、冗詞(redundant words)、選詞錯誤(incorrect word selection)、詞序錯誤(word ordering error)等語法錯誤，若能建立檢測語法錯誤的自動化系統，將有助於華語學習。該研究採用 TOCFL 語料庫，手動標記錯誤，總共有 25057 個病句和 63446 個非病句，他們以 ELECTRA (Efficiently Learning as Encoder that Classifiers Token Replacements Accurately)作為模型架構，並比較六種偵測模型的表現，分別是 CNN-LSTM、MC-CNN-BiLSTM、BERT、RoBERTa、XLNet、ELECTRA，以精確率、召回率、F1 分數作為評估指標，ELECTRA 的成績為精確率 0.6406、召回率 0.6303、F1 分數 0.6353。

2.2.2 華語病句修正

此類研究大多把語法錯誤修正視為翻譯任務，由模型判斷語料，並經過一系列的處理程序，最後輸出正確句，也就是把病句翻譯成非病句，以下為相關研究。

Ren et al. (2018)透過單詞的序列有效取得上下文的結構，他們採用卷積序列對序列模型(Convolutional Sequence to Sequence Model)，語料來源為 NLPCC 2018 國際會議提供的語料，該語料來自 Lang-8 網站，總共 1220069 個句子，先使用 Jieba 進行斷詞，再運用次字方法(subword method)來解決稀有詞和未知詞(out-of-vocabulary，簡稱 OOV)的問題，接續採用詞嵌入(word embedding)的方式處理，模型表現為精確率 0.4763、召回率 0.1256、F0.5 分數 0.3057。

Li et al. (2019) 針對華語語法錯誤(Chinese Grammatical Error Correction，簡稱 CGEC)進行修正，並提出兩個優化卷積序列對序列模型的方式，分別為共享嵌入(shared embedding)和策略梯度(policy gradient)，語料來源為 NLPCC 2018 國際會議和 NLPTEA 競賽所提供的語料，NLPCC 2018 的語料來源為 Lang-8 網站，因錯誤修正者大多是網路的華語母語者，故存在病句和非病句無法配對的狀況，有些語料為非完整修正句子，此外，NLPCC 2018 涵蓋繁體中文的語料，該研究是以簡體中文進行處理，所以需把繁體中文轉成簡體中文，接續設定條件過濾語料，避免語料無法配對的問題；NLPTEA 的語

料來自 HSK 語料庫，除了提供病句之外，還有病句錯誤的位置和類型，且病句皆由專家修改，修改的資訊幾乎沒有錯誤，語料品質高，這些都是 NLPTEA 語料庫的優點，但其缺點為語料庫太小，數量不足以處理語法錯誤偵測，有加入共享嵌入方式的模型表現為精確率 0.4333、召回率 0.0655、F0.5 分數 0.2041；有加入策略梯度方式的模型表現為精確率 0.3953、召回率 0.0746、F0.5 分數 0.2125。

Zhao & Wang (2020)指出目前已廣泛運用神經網路機器翻譯(Neural Machine Translation，簡稱 NMT)來處理語法錯誤修正此類的翻譯任務，是當前自然語言處理的應用重點，該研究把語法錯誤修正視為翻譯任務，語料來源為 NLPCC 2018 國際會議提供的語料，並使用 OpenNMT-py 作為模型架構，這是一種 PyTorch 的神經網路翻譯工具，模型表現為精確率 0.4436、召回率 0.2218、F0.5 分數 0.3697。

Wang et al. (2020)提到語法錯誤修正可視為一個序列對序列的任務，語料來源為 NLPCC 2018 國際會議提供的語料，先把所有的句子分割成漢字，其原因在於該研究的華語預訓練模型是基於字元進行處理，接著設定條件過濾語料，此外，該研究參考 HSK 語料庫的五種標記錯誤方式，分別為錯字和標點符號錯誤 B、選詞錯誤 CC、漏詞 CQ、冗詞 CD、句子錯誤 CJ，並以 Transformer 作為模型架構，目標開發華語語法錯誤修正模型，因 BERT 只使用 Transformer 的編碼器(encoder)，無法用於翻譯任務，故採用的是 Chinese-RoBERTa-wwm-ext 結合編碼器-解碼器(encoder-decoder)。從結果顯示，該研究能處理單詞層面，但句子層面仍有挑戰性，需要用其他方式解決，BERT-encoder (4-ensemble)的模型表現為精確率 0.4194、召回率 0.2202、F0.5 分數 0.3551；BERT-fused (4-ensemble)的模型表現為精確率 0.322、召回率 0.2316、F0.5 分數 0.2987。

Chen & Zhang (2022)研究有關先天聽力受損的華語學習者之語法錯誤，對於先天聽力受損的學習者而言，第一語言為手語，第二語言為書面語，兩者為不同的語言系統，若以華語作為外語的華語學習者相比，先天聽力受損的學習者更可能存在文本語法錯誤的問題。他們的語料主要來自小學至高中的聾啞人士學校的作文和日記等文本，並由語言學領域的教師和學生進行校對，並以編碼器-解碼器作為模型架構，從學習者的病句取得其中的特徵，再把句子進行重新構句，而形成符合語法的句子。透過重新排序策略，結果證實該模型可以修正多種語法錯誤。文中提到大多以 n-gram、BERT 等模型進行偵測語法錯誤，尚未完全解決詞序不當或句子層面等問題，但透過編碼器-解碼器可解決，且在一定程度上修正詞序不當的病句，其有兩個缺點，一為病句和非病句的配對數量不

足，二為不能一次修正病句的所有錯誤，甚至可能將正確的部分修改成錯誤，模型表現為精確率 0.4285、召回率 0.1635、F0.5 分數 0.3236。

Kuang et al. (2022)指出華語學習者在書寫作文時，容易發生語法錯誤的狀況，此外，網路世代興起而出現各種文本，例如新聞、部落格、電子郵件等形式，難免會出現許多語法錯誤的文本。語法錯誤修正是把病句修正成語法正確的句子，序列生成(sequence generation)和序列標註(sequence tagging)是學術界處理語法錯誤修正的方式。該研究的語料來源為 NLPCC 2018 國際會議提供的語料，並提出 CGEC-IT (Chinese GEC method based on iterative training and sequence tagging)模型，模型表現為精確率 0.362、召回率 0.213、F0.5 分數 0.318。

2.2.3 批判性回顧

總結以上有關自然語言處理於華語病句的應用，可分為病句偵測和病句修正，相關研究皆提到語法錯誤修正是目前自然語言處理所關注的議題。以往作文需藉由教師實際批改，才能判斷其錯誤，不僅花費大量人力和時間，且學習者無法即時得到反饋，也就無法即時修改錯誤。若能開發此類型的自動化語法修正系統，將有助於華語學習者練習寫作。

以華語病句偵測而言，自然語言處理領域大多把華語病句偵測視為二元分類，模型僅判斷句子是否為病句，本論文認為此部分無法完整解釋病句的語法錯誤；以華語病句修正而言，自然語言處理領域大多把華語病句修正視為翻譯任務，也就是把病句翻譯成非病句，以序列對序列為主要的處理方式，雖然這種方式能達到修正的效果，且能直接更正其錯誤，但無法提供華語病句的錯誤說明，此外，各個模型的表現結果大致落在三至五成，仍需持續增進系統的偵測及修正之能力。

除此之外，以華語系統而言，相關研究所採用的語料庫可分為兩種，一為繁體中文的臺灣華語文能力測驗 TOCFL 語料庫，語料主要是來自該測驗的寫作考試，背景為 46 種不同母語的華語學習者；二為簡體中文的中國大陸漢語水平考試 HSK 語料庫，語料來源也是來自華語學習者的寫作考試，兩者的共同點為強調以華語作為外語的概念，若模型的訓練語料來自這兩個語料庫，則較貼近華語學習者的語法錯誤，也相對適合華語學習者作為輔助學習的方式。語料量也是需考量的重點之一，若採用檢定考試類型的語

料庫，代表研究所需的語料量大，語料需達到一定的數量才適合進行相關實驗，但難點在於不容易搜集及取得如此大量的語料，且語料標註也需花費大量人力和時間。

本論文歸納以下幾點批判性回顧：

1. 把華語病句偵測視為二元分類，無法完整解釋病句的錯誤為何。
2. 把華語病句修正視為翻譯任務，無法提供華語病句的錯誤說明。
3. 自然語言處理相關研究的評估指標分數偏低。
4. 自然語言處理領域使用大型語料庫作為語料來源，代表研究所需語料量大，最關鍵的難點為語料搜集不易，以及語料標註需花費相當大的資源。

2.3 電腦輔助語言學習

在網路科技發達的世代中，各個領域積極善用電腦科技來提升工作效率，同時語言教學領域也不例外，以電腦輔助語言學習為目標，增進學習效率，已形成一個主要趨勢。

隨著電腦科技快速發展，興起一門學科稱為電腦輔助語言學習，相關研究日益廣泛（陳亮光 2008；鄭琇仁 2009；張于忻 2010；林翠雲 2012；詹衛東 2012；林翠雲 2013；陳姮良 2014；鄭琇仁 2014；詹衛東等人 2015；許德寶 2015；Lin et al. 2017；Tseng 2017；Hsin et al. 2017；Sung & Cheng 2017；李詩敏、林慶隆 2018；洪嘉馡等人 2018；連育仁 2018；Bai et al. 2019；曾妙芬等人 2019；Valdebenito & Chen 2019；Xu 2020；Tian 2020；洪嘉馡 2021；張莉萍 2022 等）。

許德寶(2015)介紹和歸納電腦輔助語言學習的名稱由來、研究發展、未來新趨勢等面向，一開始的內容先談到學科名稱，起初為電腦輔助語言教學(Computer Assisted Language Instruction，簡稱 CALI)，後來以電腦輔助語言學習(CALL)為主，兩者的差別在於後者是以學習者為中心，以雙向互動式和自主個體化為學習特點，此外，關於電腦輔助語言學習的理論框架尚未有明確定義，目前採用的研究理論分別為第二語言習得或 Bax (2003)所提出的社會學「正常化」理論，而研究方法分為語言學或社會學，最後歸納電腦輔助語言學習研究的新趨勢和未來方向。談到以學習者為教學核心，目標在於學習成效，此外，教學是一種雙向互動，關鍵角色是教師和學生。

在本節中，2.3.1 為科技應用的華語師資培訓課程，現況有越來越多科技應用的師

資培訓課程，提供教師增進科技能力，促進科技融入語言教學的發展；2.3.2 為運用多種科技工具於華語教學，善用各種科技工具，結合多元任務，目標提高學習效益，並評估其課程規劃；2.3.3 為語料庫與華語語法教學，以語料庫為語料來源，針對特定的華語語法進行語言分析；2.3.4 為多元數位華語教材，華語教材不限於傳統紙本；2.3.5 為批判性回顧，提出本論文對於電腦輔助語言學習相關研究的批判性回顧。

2.3.1 科技應用的華語師資培訓課程

科技如何結合課堂安排，對於教師而言，科技是一種新教學技能，可透過各種師資培訓課程增進其科技能力，相關研究越來越多。

陳亮光(2008)提到華語文教育以培育華語思考能力為目標，而不只是聽說讀寫等技能，以往大多以華語教材為研究方向，嘗試教材標準化等方式，效果有限，提問是教學策略之一，但教師的提問技巧普遍不理想，該研究證實華語教師應提升提問能力，並藉由多媒體科技的輔助，在教學過程中展現互動和溝通功能，進而增進華語學習成效。

鄭琇仁(2009)探討印尼地區華語師資與多媒體教學相關議題，文中提到多數華語教師認同科技輔助語言教學的優勢，但實際使用多媒體的華語教師相對較少，因此，科技應用相關的華語師資培訓課程有其重要性，該研究的受試者為印尼當地的華語教師，以問卷調查的方式得知華語師資培訓課程的成效，問卷內容主要分成三部分，一為基本資料，二為電腦使用的熟識度，三為印尼地區的多媒體華語教學相關問題，研究目的在於了解印尼華語教師的先備知識及印尼地區的多媒體華語教學狀況。

林翠雲(2013)提到實踐遠距教學的重點在於華語教師須掌握數位環境的教學設計，才有助於提升其教學效益，該研究說明遠距教學的工作流程、師資培訓課程的規劃、搭配視訊軟體的教學設計等議題，並歸納相關建議，華語教師需培養良好的帶領能力，以及非常熟悉數位工具，並鼓勵學習者多發表自己的意見，也需隨時關注學習者的狀況，以課程內容結合實際的生活經驗作為教學話題，進而從中提高學習動機。

鄭琇仁(2014)運用科技教學學科知識(Technological Pedagogical Content Knowledge，簡稱 TPACK)架構於線上華語師資的培訓課程，原先是 Shulman (1986)提出教學知識(Pedagogical Content Knowledge，簡稱 PCK)，接續由 Koehler & Mishra (2005)延伸，此架構有三個指標，分別為學門內容知識(Content Knowledge，簡稱 CK)、教學知識

(Pedagogical Knowledge，簡稱 PK)、科技知識(Technological Knowledge，簡稱 TK)，該研究有兩個研究重點，一為科技教學學科知識架構對師培生的學習啟示，二為基於科技教學學科知識架構檢測師培生的表現。

華語課程的形式不限於實體教室，已漸漸轉型為線上，目前處於實體課程和線上課程同時並行的狀態，兩者相較之下，線上課程的優點為即使人們相隔兩地，只要藉由視訊軟體，即可進行遠距華語教學，距離和時間不再是學習阻礙，方便又省時，而且，以語言學習而言，大幅增加學習者與母語者的語言交流機會，但這也代表教師需具備相關科技能力，妥善安排線上課程的運作，善用科技工具輔助教學。

林翠雲(2012)採用設計本位研究法(Designed-based Research，簡稱 DBR)，以分析(analysis)、設計(design)、發展(development)、實施(implementation)、評鑑(evaluation)的「ADDIE」流程來檢視華語教師的數位應用，該研究選用五種數位工具，分別為 VoiceThread、Voki、 GoAnimate、Penzu、Google Document，研究對象為臺灣僑務委員會開設的「99 年度華文網路種子師資培訓班」課程之「Web 2.0 工具之應用與社群經營」的學員，評估華語教師使用這些數位工具是否能達到教學目標，以及是否能提高有效學習等議題。

Hsin et al. (2017)觀察美國某高中的華語遠距教學合作項目，線上課程的教師為受過正規華語教學訓練的臺灣某大學研究生，該研究目標為提升研究生的教學技能，也讓華語學習環境更加多元。

Tseng (2017)探討線上華語教師的暑期密集培訓課程，採用混合式培訓模式，共計五週的時間，課程規劃為兩週線上培訓和三週線上試教現場指導，結合理論層面的教學方法和實務層面的電腦科技，並進行質性和量化的分析，該研究結果說明此類密集培訓課程有助於提升教師的教學能力和科技能力。

科技能力對於華語教師已是必備條件，如何規劃有關科技應用的華語師資培訓課程是一門重要的課題。Lin et al. (2017)歸納科技與華語師資培訓相關研究的發展，以及指出研究缺口，把近期的研究分為三個方向，一為科技對華語教師之教育標準，二為科技對華語教師之教育，三為華語教師使用科技的影響因素。

2.3.2 運用多種科技工具於華語教學

關於運用科技於語言教學的研究，主要是採用多種科技工具來安排課程環節，並結合多元任務，例如視訊會議、地圖導航、圖片檔案、問答遊戲、社群媒體等程式軟體，並透過一系列的教學設計，藉由學生的反饋來觀察學習成效。

陳姮良(2014)透過教學平台與翻轉模式設計華語課程，該研究認為以語言教學而言，翻轉教學的意義在於學習者先自行在家中了解語言的基本要素，接續於課堂時間，由華語教師帶領學習者進行語言練習相關活動，以提高學習動機。

Sung & Cheng (2017)探討有關一對一桌面視訊會議(Desktop Video Conferencing，簡稱 DVC)的語言教學和學習，以及教師和學生對於線上課程的反饋，研究對象為 12 位實習教師和 12 位以華語作為第二語言(Chinese as a Second Language，簡稱 CSL)的美國學習者，該研究選用 Skype 作為視訊工具，該研究結果指出即使新手教師的教學經驗有限，但他們在線上課程中採用多元的教學方式，仍能提高華語教學的成效。

洪嘉馡等人(2018)應用「華語文寫作自動評分與教學回饋平臺(Automated Essay Scoring for Han，簡稱 AES-Han)」，由國立臺灣師範大學「華語文寫作語料庫」團隊所開發，並結合「中文階層式語法庫」作為輔助工具，協助華語學習者完成自傳，目標提升華語學習者的寫作思維，並從中評估學習成效。

Bai et al. (2019)探討高級程度的華語課程結合科技輔助的學習效果，運用 Zoom、VoiceThread、語料庫等科技工具，並將課程分為個人面談、個人報告、同儕反饋、團體討論、輔導課等部分，幫助學生提高批判性思考和自主學習的能力。

語言最重要的功能在於與人溝通，如何使用語言牽涉到文化背景，可視為一種社會互動。Valdebenito & Chen (2019)提到建構主義學派認為學習者透過社會文化的互動，並經由知識內化而建構意義的過程，以達到學習效果，該研究基於此概念，課程目標為華語教學課堂結合科技應用，進而加強華語能力，課程主題為「食物與文化」，並搭配三種科技工具，分別為 Google MyMaps、WordPress、Adobe Spark，探討學習過程的自主性和真實性。

曾妙芬等人(2019)提到如何藉由科技工具來提高溝通效率，讓學生能夠最大化語言輸出是研究重點，在他們的研究中，介紹各種科技工具，例如 Canvas、Quizlet、PlayPosit、Zoom、Nearpod、Padlet、Flipgrid、Facebook，使用這些工具來增強線上課程的教學互動

性，並說明科技輔助確實能提高學習效率。

機器翻譯(Machine Translation，簡稱 MT)廣泛運用在我們的日常生活裡，無論是自行輸入內容或圖片辨識後轉譯成文字，且隨著時間發展，機器翻譯的能力日漸強大，相關研究主要是以 Google Translate 和其他機器翻譯系統進行比較，並設法融入課程內容，藉由電腦科技達到學習的自主性。

Xu (2020)觀察 12 位美國某大學的華語學習者，藉由機器翻譯的協助，編輯自己的作文，並以問卷和訪談的方式，從中搜集學習者的反饋，從中發現學習者認為機器翻譯能夠幫助學習，以及提升作文的質量，該研究結果顯示 70%的學習者表示機器翻譯是有用的學習工具，而其中 69%表示未來仍會繼續使用，另一方面的學習者則認為這是一種作弊的行為，從這些學習者的反思中，提到機器翻譯的缺點，部分認為如果過度使用機器翻譯，可能對自己的學習成效沒有信心，以及不認真的學習者仍可藉由機器翻譯而得到分數。

Tian (2020)設計一項華語教學任務，但研究結果不如預期，預設目標為華語學習者藉由機器翻譯的輔助，以提升寫作能力，華語學習者以華語進行寫作練習，再輸入搜狗翻譯(Sogou Translate)進行翻譯，以翻譯後的英文結果，反推評估各自的華語寫作是否有誤，華語學習者可自行修改，換言之，如果英文結果有誤，則可能是華語原句有誤，該研究結果發現機器翻譯具有一定程度的容錯性，即使華語學習者的作文有各種錯誤，但使用機器翻譯後，仍可得到可接受的英文翻譯結果，這部分也說明目前的機器翻譯相當先進。

2.3.3 語料庫與華語語法教學

語料庫可視為科技工具之一，許多語法教學的研究，針對特定的語法進行語言分析，語料主要來自語料庫，其原因在於以人工搜集語料，需花費大量人力和時間，也可能發生選取依據不一致的狀況，且語料庫的規模日漸完善，語料的數量和類別相對齊全，讓使用者方便又快速搜集教學或研究所需的素材。

詹衛東(2012)提到樹庫(Treebank)是自然語言處理的研究方向，基於句法結構，不僅是處理語言信息相關技術，也能提供語言教學作為參考，文中簡介北京大學樹庫，原始語料先經過斷句及斷詞等處理程序，再由人工使用樹庫的輔助編輯程式(Tree Editor)進

行逐句檢查，語料來源為語文課本、句型例句、新聞等，總共 55742 句，以及探討以往關注樹庫的自然語言處理之應用，研究目的在於提供華語教師和學習者能透過樹庫系統查詢句型和例句，並作為一個句型練習平台。

詹衛東等人(2015)以華語述補結構為研究重點，開發相關語料庫，述補結構由兩個謂詞成分組成，文中舉例吃飽、洗乾淨等，此結構須有因果關係，其他語言需兩個子句或其他複雜結構才能表達其語意，而華語只需簡短數個字，因此，華語學習者較難掌握，且可能避免使用，研究目的在於經由一系列的處理程序，提出一套具可行性的計算公式，算出兩者之間的事件相關性，也提供一種華語教學的輔助工具。

李詩敏、林慶隆(2018)探討「高興」的類近義詞，文中指出近義詞被當作第二語言習得的掌握關鍵，近義詞的詞彙意義很接近，但用法不能完全替換，仍有些許限制和差異，導致學習者難以辨別，以及使用時容易混淆。該研究採用國家教育研究院的語義場關聯詞查詢系統和華英雙語索引典系統，並搭配國立臺灣師範大學的雙語索引典，其原因在於若請專家找尋辭典的近義詞，可能發生選詞規則不夠一致的狀況，且辭典的近義詞解釋常有循環解釋的問題，當近義詞數量很多時，需藉由人工檢查語料，反而花費更多人力和時間。研究目的在於偵測和找出近義詞，最後分析和歸納結果，可應用於教材編輯，以及提供教師和學生作為學習使用。

張莉萍(2022)研究有關中高級華語教學的語法點，語法知識對於學習者的語言程度至關重要，目標探討語法點的定義及各種教材版本的語法點不同之處，並提出語法點是否為華語學習者須學習的重點，基本假設為高頻率使用的語言成分，將有助於學習者習得目標語，主要是採用國家教育研究院語料庫索引典系統，搭配統計的方式，得到相關數據結果。

2.3.4 多元數位華語教材

為了確保華語學習者能全方位學習，善用科技輔助，以提高聽說讀寫技能，因此，數位華語教材資源越來越多。

張于忻(2010)提到已有明確規範數位教材的範疇，透過各面向的標準進行檢核，確保數位教材有助於學習成效。該研究以「教學設計」面向為主軸，並基於模組化方式探討如何設計華語教材。

連育仁(2018)以智慧型手機強化語言學習(Smartphone Enhanced Language Learning，簡稱 SELL)為核心概念，自行發展相關系統作為華語教材工具，探討華語教師對於行動裝置整合語言教材作為數位化華語教材的接受度，並分析數位教材的教學策略，進而評估其華語教學成效。

洪嘉馡(2021)基於語料庫語言學的研究方法結合數位科技，從母語語料、學習者偏誤、教材此三個面向，建立「中文語法數位平台」，應用教學平台輔助華語教學，內容提到華語學習者目標達到語言溝通的效果，除了累積大量單詞，完整的語意仍以語法結構為基礎，因此，華語語法教學有其重要性，進而影響教學語法(pedagogical grammar)受到重視，以句子種類、句子架構、句子主要成分作為「中文階層式語法點」，並評估該教學平台對於華語教學的效益，該教學平台基於數位中文語法庫，並整合多元功能，善用語料庫來保存相關語料，教師能從中節省時間來查詢資料，學習者也能不受時間限制進行學習，皆有助於提升華語學習的成效。

2.3.5 批判性回顧

上述提到多種電腦科技輔助華語教學和學習的面向，從職前科技應用的師資培訓課程，提升教師的科技能力，接續運用電腦科技於實際華語教學課堂，善用多種科技工具，設計多元任務，提高學習動機，可以看出每個環節皆致力於電腦科技結合華語教學。

相關研究皆證實電腦科技確實有助於語言學習，雖然 Tian (2020)的研究結果不如預期，但從中發現機器翻譯有一定程度的容錯性，這也說明目前機器翻譯的能力相當強大，因此，並非機器翻譯不適合應用於語言教學，而是需考量更多教學細節，並納入課程規劃之中。

語料庫也可視為一種科技工具，許多語法教學相關研究都以語料庫作為語料來源，搜集語料是一個很繁瑣的過程，採用語料庫可避免耗費過多人力和時間的阻礙，此外，開發華語相關語料庫，目標提供華語教師一種教材檢索系統，同時也是華語學習者的線上練習平台，但語料庫相關研究非以偵測及修正華語病句為重點。

以電腦輔助語言學習領域而言，大多以華語師資和教學課程為主軸，例如規劃科技應用的師資培訓課程，以及採用多種科技工具輔助華語教學，相較之下，有關華語病句的研究較少見，但華語教師若能熟悉學習者的偏誤狀況，將有助於累積華語教學的實務

經驗，故華語病句相關研究有其重要性。

再者，機器翻譯是一種很方便的工具，學習者能使用機器翻譯來輔助練習寫作，但大多機器翻譯通用多種語言，並非以華語為主要設計，有些語句不一定完全符合實際用法，此外，語料庫收錄各種類型的素材，讓使用者方便搜集語料，可避免花費大量人力和時間的問題，採用語料庫的研究大多針對特定的語法進行語言分析，較少看到整合式的系統，所以無法提供使用者客製化的語法說明。

本論文歸納以下幾點批判性回顧：

1. 目前電腦輔助語言學習領域有關華語病句的研究較少見。
2. 語料庫相關研究非以偵測及修正華語病句為重點，且較少看到整合式的系統，故無法提供使用者客製化的語法說明。

2.4 結語

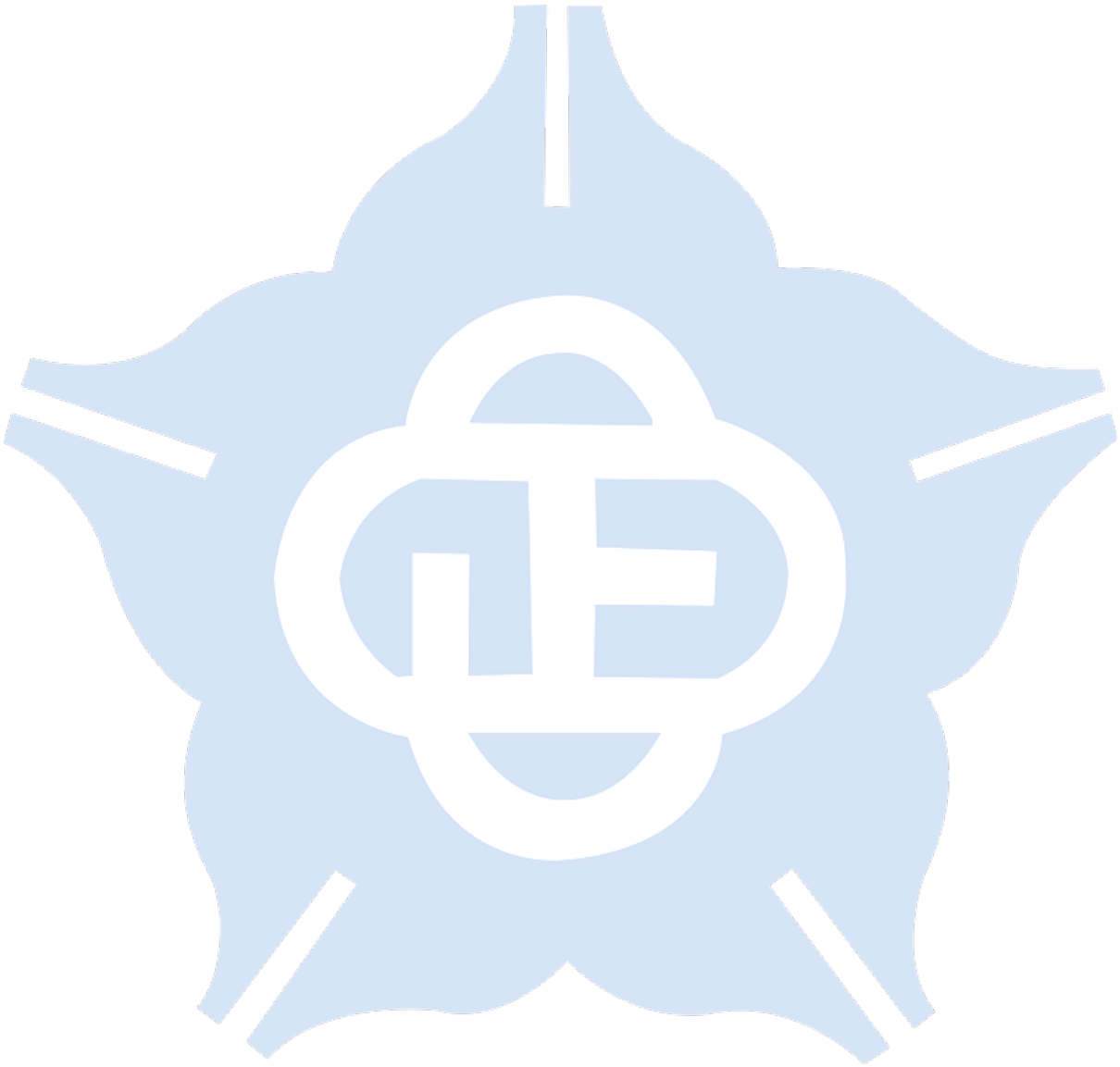
本章回顧自然語言處理領域和電腦輔助語言學習領域的文獻，華語對於這兩大領域皆是重要議題。自然語言處理領域使用不同的模型來進行病句偵測和病句修正之任務，而電腦輔助語言學習領域強調華語教學結合科技應用，各研究皆致力於善用科技工具來輔助語言教學和學習。

關於自然語言處理技術於華語病句的應用，研究方向可分成病句偵測和病句修正。以病句偵測而言，大多以二元分類進行處理，模型僅判斷句子是否為病句，但無法完整解釋病句的錯誤為何；以病句修正而言，自然語言處理領域大多視為翻譯任務，以序列對序列為主要的處理方式，雖然這種方式能達到修正效果，且能直接更正其錯誤，但無法提供華語病句的錯誤說明。除此之外，相關研究主要採用華語能力相關檢定的語料庫作為語料來源，代表研究所需的語料量大，最關鍵的難題在於語料搜集不易，以及語料標註需花費大量人力和時間。

電腦輔助語言學習相關研究大致可分成華語師資培訓、華語教學課堂、華語教材等主題，但華語病句相關研究較少見，熟悉華語學習者的偏誤狀況有助於累積華語教學的實務經驗，因此，華語病句相關研究有其重要性。

本論文以華語病句為研究重點，整合華語教學與自然語言處理技術，把華語病句視

為分類任務，由模型經過三種分類，依序為病句判斷、錯誤主類別判斷、錯誤次類別判斷，逐步確認病句的錯誤分類，最後提供病句的建議說法和錯誤說明，藉此華語學習者能從中初步了解錯誤原因，將有助於實踐電腦輔助語言學習之目標。



第三章 語料說明及實驗方法

3.1 概述

自然語言處理領域大多把華語病句修正視為翻譯任務，但本論文則視為分類任務。本論文選用三種模型，基於語言學理論的 LOKI、統計式簡單貝氏分類器、深度學習神經網路 LSTM，針對相同數量的少量語料進行參照比較。

本實驗分成訓練階段和測試階段，並在每個階段中，模型均進行三種分類，依序為病句判斷、錯誤主類別判斷、錯誤次類別判斷，逐步確認病句的錯誤分類，才能提供華語病句的建議說法和錯誤說明。在相同數量的少量語料下，比較此三種模型的判斷結果，並計算正確率、精確率、召回率、F1 分數此四個成績作為評估指標。

本章架構如下：3.2 為語料說明，介紹語料來源、語料處理程序等內容；3.3 為實驗說明，介紹 LOKI、簡單貝氏分類器、LSTM 的運作原理，以及說明各項評估指標；3.4 為實驗步驟，介紹本實驗的處理程序；3.5 為訓練階段，說明三種模型的訓練過程及相關模組；3.6.為測試階段，說明三種模型的測試過程及相關模組；3.7 為結語，總結本章內容。

3.2 語料說明

本論文的語料庫總共有 640 個句子，訓練階段的語料庫共有 540 個句子，分別為 340 個病句和 200 個非病句，病句語料來自美國某大學華語領航學程的課堂寫作，而非病句語料來自國家教育研究院的華語中介語索引典系統；測試階段的語料庫共有 100 個句子，分別為 50 個病句和 50 個非病句，語料來自國家教育研究院的華語中介語索引典系統。

在本節中，3.2.1 為語料來源，詳細說明本論文的語料來源；3.2.2 為華語學習者的學習背景和程度，介紹華語學習者的學習背景及作文程度；3.2.3 為語料處理程序，把已整理的作文語料由簡體字轉成繁體字，從中搜集病句，接續進行病句斷詞等處理程序。

3.2.1 語料來源

關於訓練階段的語料庫，病句語料來自美國某大學華語領航學程的課堂寫作，有 35 位華語學習者，總共採用 47 篇華語寫作文章。寫作主題豐富，例如日常生活、大學回憶、飲食文化、閱讀心得、現況時事等。

這些語料先由在美國任教多年的專業資深華語教師進行寫作分級，依照美國外語教學委員會(American Council on the Teaching of Foreign Languages，簡稱 ACTFL)的語言能力架構，華語學習者皆來自美國某大學華語領航學程，其華語程度相對較高，他們的作文程度大致落在中級以上。華語的句子單獨出現及出現於篇章有不同的表現，例如句子是否符合語法、語意是否完整等，因中級程度以上的華語學習者較能寫出相對完整的華語篇章，從中可看出華語學習者是否能正確使用華語、掌握華語的用法等現象，故本論文採用中級程度以上的華語寫作語料，並從中搜集 340 個華語病句。

非病句語料來自國家教育研究院華語文語料庫與能力基準整合應用系統之華語中介語索引典系統，從中採用 200 個非病句，選用此語料庫的原因有三，一為本論文考量原始語料的句數不足，無法提供足夠的非病句，故需另外採用其他語料庫；二為本論文設定的語料來源盡可能以華語學習者為主，而非華語母語者，該語料庫主要收錄華語學習者的寫作文章，其語料同時兼具病句和非病句；三為該語料庫的作文篇幅較完整，而非只是單句，因此，推論這些作文的華語學習者也具有一定程度的寫作能力。

有關偏誤分析的研究，內容可能會提到不同程度的華語學習者有共同常見的偏誤狀況，但本論文不討論華語病句分級相關議題，故不處理此議題。整體而言，本論文的語料庫共有 640 個句子，其中 540 個句子作為訓練階段的語料庫，從中再分成訓練集和驗證集，以及其中 100 個句子作為測試集，主要用於測試階段，分別為病句和非病句各 50 個，把測試集輸入至已訓練的模型進行預測，並計算評估指標作為模型的測試結果，測試集的語料來源如同上述的國家教育研究院華語文語料庫與能力基準整合應用系統之華語中介語索引典系統。

3.2.2 華語學習者的學習背景和程度

本論文採用四個班級的語料，每個班的程度不同，班級屬性分為領航班和普通班，其中三個班為領航班，另一個班為普通班，領航班的學生程度相對較高，為了擴大語料範圍，了解不同程度的華語學習者之偏誤狀況，故採用兩種不同屬性的班級作為病句的語料來源。領航班為華語學習者申請領航課程項目，順利通過而成為領航班的成員，課程進度較密集，且定期都會舉辦考核，檢視學習者的學習狀況，有些學習者可能不符合項目要求，或有其他安排而無法繼續待在領航課程，則選擇普通班繼續學習華語；普通班類似大學的通識課程，而該華語課是屬於語言課程，而不是文學課程。以學習者的程度而言，普通班的學習者非較差，但普通課程的進度不像領航課程如此緊湊，因此，長期累積下來，學習者的程度仍是有所分別。

以本論文的語料而言，華語學習者的課堂教材皆來自國立臺灣師範大學國語教學中心編寫的《當代中文課程》系列，整體而言，四個班級的華語程度大致落在教材第三冊至第六冊之間，作文涵蓋至少中級以上的程度，語料特點為寫作主題豐富、用詞生活化、句型較多元等，因語料來自實際的美國大學華語課堂，故病句的錯誤類型可視為以英語為母語的華語學習者之偏誤狀況。

3.2.3 語料處理程序

本論文的病句語料來自課堂寫作，這些語料先由在美國任教多年的專業資深華語教師進行寫作分級，採用的是 ACTFL 語言能力架構。ACTFL 總共有十一個分級，此架構的特點在於每個等級涵蓋較低等級的程度，且針對不同的語言而設計相關能力描述，是當前分級最詳細的語言評估架構(熊玉雯等人，2014)。ACTFL 分級圖如圖 1，圖片來自 ACTFL 網站。[1]

[1] ACTFL 網站：https://www.actfl.org/educator-resources/actfl-proficiency-guidelines

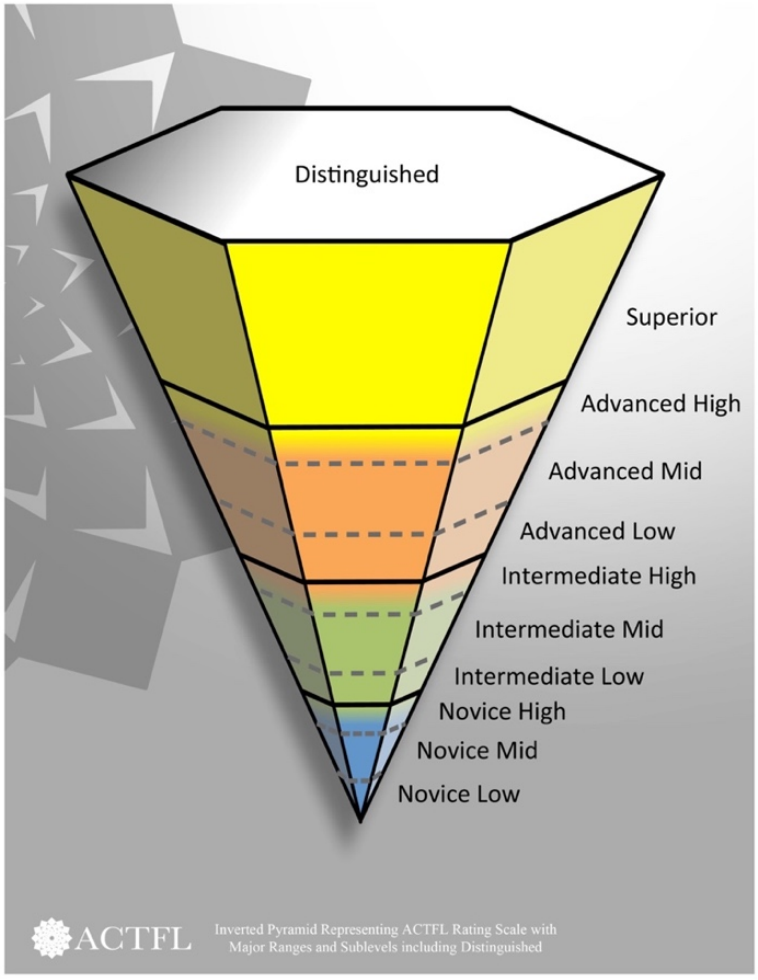


圖 1. ACTFL 分級圖

本論文的華語學習者以華語領航學程的學生為主，其華語程度相對較高，這些作文語料涵蓋至少中級以上的程度，從中級中等(Intermediate Mid)至高級高等(Advanced High)之間。中級的作文篇數為 15 篇，高級的作文篇數為 32 篇，兩者共計 47 篇。以各級病句數量而言，中級的病句數量為 134 個，而高級的病句數量為 206 個，兩者共計 340 個病句。

華語學習者以簡體字書寫作文，本論文先把全部的作文語料轉成繁體字，再以人工判斷句子是否有錯誤，如果句子有至少一個錯誤，即視為病句，本論文總共搜集 340 個病句作為病句語料庫。

華語病句為不符合華語語法的句子，觀察病句語料的偏誤狀況，可分成三種錯誤主類別，分別為語法(syntax)、語意(semantics)、詞彙(vocabulary)。語法錯誤為句子結構有

誤，例如過度類化使用「了」、副詞「也」的位置錯誤、分裂句等；語意錯誤為句子的語意不明確或有歧義的狀況，需再次提問來確認語意，例如「他們有個電影晚上」，語意可能是「他們在晚上看電影」或「他們要去看午夜場電影」等；詞彙錯誤為句子中包括至少一個詞彙使用錯誤，例如近義詞、音調問題、同音字等，此外，錯字僅需修改該字，以及重複字詞僅需去除多餘的字詞，所以錯字及重複字詞皆屬於詞彙錯誤。各種錯誤主類別的病句數量為語法 123 個、語意 69 個、詞彙 148 個，資料統整如表 1。

因語意的錯誤顯現在句型結構上，且詞類是句型結構的成分之一，例如「春草發現了她被騙」，此病句的錯誤在於過度類化使用「了」，華語學習者皆為英語母語者，把華語「了」當成英語過去式的標記，華語學習者藉由「了」來表達過去發生的事情，但動詞「發現」後不需加「了」。關於動詞後可否接完成貌「了」，這是一個語意問題，有相關研究探討此議題，例如 Huang (2009)等，但不同動詞在可否接完成貌「了」的表現有所差異，故可從詞類來辨識其呈現方式。由此可見，詞類也是影響句型結構的因素，且詞彙的構詞方式與句型結構的概念相關，此外，語法規則是基於句型結構，故本論文把每個病句代表一種錯誤句型，所以本論文共有 340 個錯誤句型作為錯誤次類別。

錯誤主類別	判斷依據	數量
語法(syntax)	句子結構有誤	123
語意(semantics)	句子的語意不明確或有歧義的狀況	69
詞彙(vocabulary)	詞彙使用錯誤、錯字、重複字詞	148
共計		340

表 1. 錯誤主類別的病句數量

整理完華語病句後，以卓騰語言科技公司開發的文截斷詞系統(Wang et al. 2019，Articut)進行病句斷詞，為了維持實驗標準的一致性，因 LOKI 的斷詞系統為 Articut，故採用相同的斷詞系統。

接續為非病句語料，依照語法、語意、詞彙此三種錯誤主類別來搜集，例如華語學習者誤用同音字「在」和「再」，基於這組同音字，本論文從語料庫中搜尋正確的華語用法作為非病句語料，非病句語料來自國家教育研究院華語文語料庫與能力基準整合應

用系統之華語中介語索引典系統，從中採用 200 個非病句，接續進行非病句斷詞。

最後有關測試階段的語料，同樣來自國家教育研究院華語文語料庫與能力基準整合應用系統之華語中介語索引典系統，本論文隨機選取及搜集 100 個句子作為測試集，分別為病句和非病句各 50 個，接續進行測試集斷詞。

本論文統整語料處理程序：

1. 搜集華語課堂的寫作文章。
2. 經由專業資深的華語教師進行華語寫作分級。
3. 把簡體字轉為繁體字，並從中搜集病句語料。
4. 病句錯誤分類，錯誤主類別分成語法、語意、詞彙。
5. 採用 Articut 進行病句斷詞。
6. 基於錯誤主類別來搜尋非病句語料。
7. 採用 Articut 進行非病句斷詞。
8. 隨機選取及搜集測試階段的語料。
9. 採用 Articut 進行測試集斷詞。

3.3 實驗說明

本論文探討在相同數量的少量語料下，比較 LOKI、簡單貝氏分類器、LSTM 此三種模型的表現差異。本節介紹此三種模型的基本原理和運作模式，以及說明正確率、精確率、召回率、F1 分數等各項評估指標。

在本節中，3.3.1 為語言導向的關鍵詞介面(LOKI)，介紹 LOKI 的運作原理；3.3.2 為統計式簡單貝氏分類器(Naïve Bayesian Classifier)，介紹簡單貝氏分類器的運作原理；3.3.3 為深度學習神經網路長短程記憶模型(LSTM)，介紹 LSTM 的運作原理；3.3.4 為評估指標說明，解釋正確率、精確率、召回率、F1 分數此四個評估指標的意義。

3.3.1 語言導向的關鍵詞介面(LOKI)

LOKI 是由卓騰語言科技公司開發的產品，這是一種自然語言理解(Natural Language Understanding，簡稱 NLU)的引擎。LOKI 基於語言學理論為運作原理，其強大的特色在於以一個句子即可辨識相同句型的多個句子。

首先，有關句子結構，句子最基本的單位為單詞(word)，每個單詞的詞類決定其於句子中的位置，因此，在分析句子層面時，詞類被視為相當重要的元素。根據卓騰語言科技公司的實務經驗，結合語法和語意的概念，把詞類分成七大類，分別為實體類(ENTITY)、動詞類(ACTION、ASPECT、MODAL、AUX)、時間類(TIME)、修飾詞(IDIOM、MODIFIER、MODIFIER_color、ModifierP、DegreeP、QUANTIFIER)、功能詞(FUNC)、句型詞(CLAUSE)、NER 類(LOCATION、KNOWLEDGE、UserDefined、RANGE)，資料統整如表 2，表格內容來自卓騰語言科技公司網站。[2]

詞類	細項
實體類	ENTITY
動詞類	ACTION、ASPECT、MODAL、AUX
時間類	TIME
修飾詞	IDIOM、MODIFIER、MODIFIER_color、ModifierP、DegreeP、QUANTIFIER
功能詞	FUNC
句型詞	CLAUSE
NER 類	LOCATION、KNOWLEDGE、UserDefined、RANGE

表 2. Articut 詞類

[2] 卓騰語言科技公司網站：https://api.droidtown.co/document/#Pos

LOKI 的運作原理基於語言學的句法分析，以動詞作為句型結構的中心，並以詞組結構律(Phrase Structure Rules)為核心概念。相同詞類的單詞可出現在句子中的相同位置，所以在斷詞及詞類標註後，就能用相同詞類的單詞來取代原本訓練語料的單詞，藉此達到以一個句子即可辨識相同句型的多個句子之目標。

在此以「我住在臺灣」為例句，經由 Articut 進行斷詞，其結果為「我/住/在/臺灣」，例句的詞類為「ENTITY_pronoun / ACTION_verb / FUNC_inner / LOCATION」，其中的代名詞「我」為可有可無(optional)的元素；[3]因動詞是必要元素，故動詞「住」不可更換成其他相同詞類的動詞，以此例句的句型而言，動詞的位置必須為「住」；功能詞「在」被 LOKI 視為虛詞，所以不做處理；地點「臺灣」也是必要元素，但可替換其他相同詞類的單詞，例如其他國家或地名。簡而言之，除了動詞以外，其他詞類的位置可替換成相同詞類的單詞，因此，「她住在美國」、「住在日本」這類的句子皆可對上此句型。

基於詞組結構律的運作概念，其強大的特色在於以一個句子即可辨識相同句型的多個句子。LOKI 是以語言學的句法分析為基礎的自然語言理解引擎，掌握句型的關鍵，實踐以少量語料即可完成辨識大量句子之目標。

3.3.2 統計式簡單貝氏分類器(Naïve Bayesian Classifier)

統計式機器學習有多種分類器，簡單貝氏分類器是其中一種，主要處理分類任務，此分類器基於貝氏定理(Bayesian Theorem)的條件機率，假設各類別是獨立不相關，很多文獻探討這種分類器的應用，例如 Leung (2007)、Rish (2001)等，且很多統計學教科書也都會提及貝氏定理，例如林惠玲、陳正倉(2018，頁 157-160)等。

林惠玲、陳正倉(2018，頁 157)提到：「貝氏定理 […] 即是說明如何由新資訊修正事前機率而得事後機率的方式。」他們以如下的舉例來解釋貝氏定理。假設有一個手機公司要推出一款新手機，公司知道之前推出手機的成功與失敗的機率：成功機率(A_1)為 0.7，失敗機率(A_2)為 0.3，因此，條件機率的格式為 $P(A_1) = 0.7$ 和 $P(A_2) = 0.3$。

根據過去經驗，在推出成功的情況之下，0.9 的機率為消費者喜歡，而 0.1 的機率

[3] 但若以句法學的看法而言，代名詞常出現於論元的位置，是必須存在的成分，而修飾詞屬於非必要的成分。

為消費者不喜歡。在推出失敗的情況之下，0.7 的機率為消費者不喜歡，而 0.3 的機率為消費者喜歡，此舉例的條件機率，資料統整如表 3。

	推出成功(A_1)	推出失敗(A_2)
消費者喜歡(B_1)	$P(B_1|A_1) = 0.9$	$P(B_1|A_2) = 0.3$
消費者不喜歡(B_2)	$P(B_2|A_1) = 0.1$	$P(B_2|A_2) = 0.7$
合計	1.0	1.0

表 3. 事前條件機率

如上述的條件機率，是否能得知如果消費者喜歡而推出成功，也就是 $P(A_1|B_1)$的機率為多少？貝氏定理的公式為 $P(A_i|B) = \frac{P(B \cap A_i)}{P(B)}$。是否能從 $P(B_1|A_1)$得到聯合機率 $P(B_1 \cap A_1)$？答案為可以，$P(B_1 \cap A_1) = P(A_1) \times P(B_1|A_1)$，聯合機率 $P(B_1 \cap A_1) = 0.7 \times 0.9 = 0.63$，在此也需得知 $P(B_1)$，$P(B_1) = P(B_1 \cap A_1) + P(B_1 \cap A_2)$，而 $P(B_1 \cap A_2) = P(A_2) \times P(B_1|A_2) = 0.3 \times 0.3 = 0.09$，因此，$P(B_1) = 0.63 + 0.09 = 0.72$。如果消費者喜歡的話，推出成功的機率為 $P(A_1|B_1) = \frac{P(B_1 \cap A_1)}{P(B_1)} = \frac{0.63}{0.72} = 0.875$，也就是如果消費者喜歡的話，推出的成功機率是 0.875。

如何運用貝氏定理來處理分類任務？以本論文的病句概念進行說明，假設我們有 100 個華語學習者所寫的句子，而每個句子都有病句或非病句的標籤。如果是病句，第一句出現的機率為何？如果是非病句，第二句出現的機率為何？後續的句子以此類推，我們可以利用上述的條件機率，從推出成功的機率、已知推出成功而喜歡的機率、喜歡的機率來得到已知喜歡而推出成功的機率，換言之，從已知某標籤在哪些句子所出現的機率來推算新句子屬於某標籤的機率。

3.3.3 深度學習神經網路長短程記憶模型(LSTM)

LSTM 是深度學習神經網路的分支，由 Hochreiter & Schmidhuber (1997)提出。較早出現的深度學習神經網路，例如 CNN 或 RNN，其缺點在於不能模擬捕捉遠距關係，而

LSTM 改善了此缺點。

在自然語言中，常有遠距關係的語言現象，句法研究的遠距依存(long-distance dependence)是一個很好的例子，例如 Chomsky (1957)、Harris (1945)等。因能夠模擬遠距依存的關係，讓較早進入神經網路處理的訊息才能保留，在處理後續的訊息時，也能讀取使用較早的訊息。LSTM 基於此優勢，故在許多自然語言處理任務的表現比 CNN 和 RNN 更好。

神經網路可視為模擬人類腦部的運作方式，許多神經元組合成一組神經網路，神經元收到訊號時，基於訊號的強弱決定神經元是否擊發，將訊號傳給下個神經元。只要能夠把訊號傳下去，則這個神經網路就可以學習。

在使用深度學習神經網路來模擬人類腦部的運作時，一切皆由數學決定，決定神經元是否擊發為激勵函數(activation function)，神經網路使用損失函數(loss function)來運算模型的計算結果與真實結果之間的誤差，並使用優化函數(optimization function)來縮小損失函數的數值，讓模型的表現能達到最佳狀態。

LSTM 有一個貫穿整個任務處理的每個時間點之記憶狀態，此部分有別於其他深度學習神經網路，這個記憶狀態讓 LSTM 在處理遠距依存關係時，其表現超越其他模型，Lane et al. (2019，頁 277)使用如圖 2 來展示 LSTM 的記憶狀態。

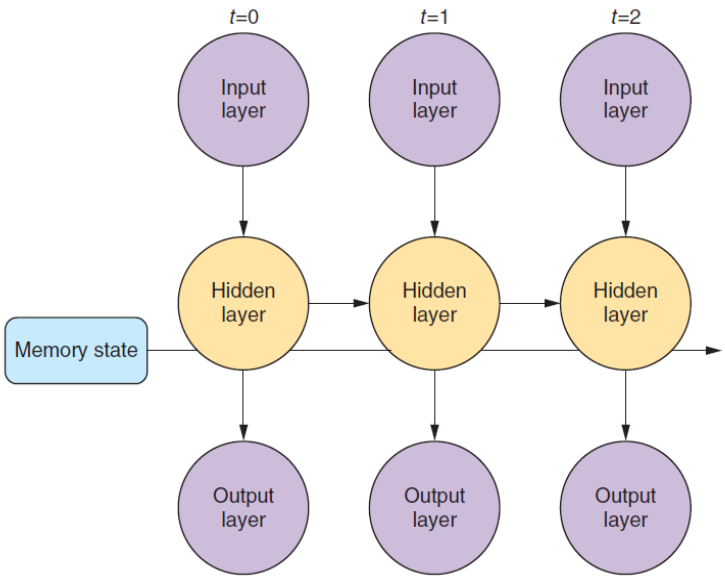


Figure 9.2 Unrolled LSTM network and its memory

圖 2. LSTM 記憶狀態圖

在上方的圖中，等待處理的字符逐一依序輸入神經網路做處理，也就是圖中所示，不同的時間點(*t*)的輸入(input)。從圖中可看到有一個記憶狀態(memory state)貫穿了整個文本的隱藏層，而這個記憶狀態能讓 LSTM 在處理一個文本的任何字符時，都能夠讀取，也就使遠距依存關係得以保留及處理。

記憶狀態也是一個前饋神經網路(forward-feeding neural network)，但如果把整個文本的所有訊息都儲存在記憶狀態中，記憶狀態會過於龐大，而形成梯度爆炸的問題，在電腦模擬時，可能造成影響計算速度或電腦記憶體無法負荷等問題。為了解決此問題，記憶狀態對已傳入的先前訊息做處理，Lane et al. (2019，頁 278)使用如圖 3 進行說明。

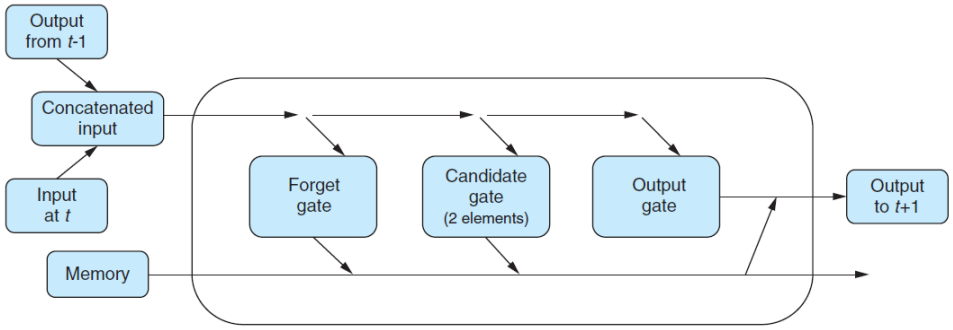


Figure 9.3 LSTM layer at time step *t*

圖 3. LSTM 記憶程序圖

如上圖所示，記憶狀態涵蓋三個閘門，分別為遺忘閘(forget gate)、候選閘(candidate gate)、輸出閘(output gate)。遺忘閘的功能為決定先前資訊是否要從記憶狀態遺忘；候選閘的功能決定先前資訊是否要記錄下來；輸出閘的功能則決定某種資訊是否要傳入隱藏層。

在本論文的研究中，模型判斷分成三種，判斷句子是否為病句、病句屬於哪種錯誤主類別、病句屬於哪種錯誤次類別，這些都是分類任務，可選用 LSTM 來完成分類任務。LSTM 除了運用於分類任務，也常搭配其他神經網路組合使用，用來執行文本生成或翻譯任務。

3.3.4 評估指標說明

關於評估機器學習的表現，大多以正確率、精確率、召回率、F1 分數作為評估指標，這些成績基於混淆矩陣(Confusion Matrix)的概念，可分為陽性(True Positive，簡稱 TP)、陰性(True Negative，簡稱 TN)、偽陽性(False Positive，簡稱 FP)、偽陰性(False Negative，簡稱 FN)此四個數值。

以本論文的病句概念進行說明，TP 為模型判斷為病句且確實為病句；TN 為模型判斷為非病句且確實為非病句；FP 為模型判斷為病句但實際是非病句；FN 為模型判斷非病句但實際是病句，換言之，TP 和 TN 是模型正確判斷的狀況，而 FP 和 FN 是模型誤判的狀況，資料統整如下方表 4。

	預測為病句(Positive)	預測為非病句(Negative)
實際為真(True)	TP	TN
實際為假(False)	FP	FN

表 4. 混淆矩陣

基於上述的 TP、TN、FP、FN 此四個數值，分別計算正確率、精確率、召回率、F1 分數，各項評估指標的計算公式如下所示。

$$\text{正確率(Accuracy)} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{精確率(Precision)} = \frac{TP}{TP+FP}$$

$$\text{召回率(Recall)} = \frac{TP}{TP+FN}$$

$$\text{F1 分數(F1-score)} = \frac{2*\text{Precision}*Recall}{\text{Precision}+Recall}$$

關於各項評估指標的重點，以本論文的病句概念進行說明。正確率為在整體模型的判斷結果下，其中正確判斷所佔的比例；精確率為在模型判斷為病句的狀況下，其中正確判斷為病句所佔的比例，其著重於 FP 的影響；召回率為在確實為病句的狀況下，其中正確判斷為病句所佔的比例，其著重於 FN 的影響；F1 分數為精確率和召回率的調和平均數，兩者均等重要。

3.4 實驗步驟

本論文把華語病句修正視為分類任務，本實驗分成兩個階段，依序為訓練階段和測試階段，訓練階段分成訓練模型和驗證模型，完成訓練和驗證的實驗步驟，接續進入測試階段。在每個階段中，模型均進行三種分類，依序為病句判斷、錯誤主類別判斷、錯誤次類別判斷，逐步確認華語病句的錯誤分類，才能提供華語病句的建議說法和錯誤說明。本論文的語料數量共有 640 個句子作為語料庫，其中 540 個句子用於訓練階段，以及其中 100 個句子用於測試階段。

關於模型判斷的三種分類，第一種分類為病句判斷，把語料輸入至模型，由模型判斷句子是否為病句，確認句子為病句後，接續第二種分類為錯誤主類別判斷，由模型判斷病句屬於哪種錯誤主類別，最後第三種分類為錯誤次類別判斷，由模型判斷病句屬於哪種錯誤次類別，逐步確認病句的錯誤分類，才能提供華語病句的建議說法和錯誤說明。

第一個階段為訓練，此階段分成訓練模型和驗證模型。第一種分類為病句判斷，屬於二元分類，共有 540 個句子作為訓練階段的語料庫，分別為 340 個病句和 200 個非病句，並依照自然語言處理的常用做法，分成 80%語料進行訓練，共有 432 個句子作為訓練集，而 20%語料進行驗證，共有 108 個句子作為驗證集。第二種分類為錯誤主類別判斷，以及第三種分類為錯誤次類別判斷，兩者屬於多元分類，為了避免第一種分類病句判斷的結果影響後續的模型判斷，因假設病句判斷有誤判的狀況，後續的模型判斷可能也會出錯，故在錯誤主類別判斷及錯誤次類別判斷上，只採用 340 個病句，語料分配同樣依照自然語言處理的常用做法，80%語料進行訓練，共有 272 個病句作為訓練集，而 20%語料進行驗證，共有 68 個病句作為驗證集，並計算正確率、精確率、召回率、F1 分數此四個成績作為評估指標，並從中比較三種模型的訓練表現。

設定語料比例的用意在於若沒有新語料作為測試集，仍可用驗證集來取得評估指標，

並作為評估模型的訓練表現之參考，但若有新語料作為測試集，就能進行模型預測，以維持實驗標準的完整性。

第二個階段為測試，本論文另外搜集 100 個句子作為測試集，其中分別為 50 個病句和 50 個非病句。在測試階段，如同上述的處理程序，LOKI、簡單貝氏分類器、LSTM 分別依序進行病句判斷、錯誤主類別判斷、錯誤次類別判斷，並計算正確率、精確率、召回率、F1 分數此四個成績作為評估指標，並從中比較三種模型的測試結果。

3.5 訓練階段

在訓練階段，本論文的語料庫分為 340 個病句和 200 個非病句，並依照自然語言處理的常用做法，分成 80%語料作為訓練集，而 20%語料作為驗證集，用驗證集進行模型預測，並計算評估指標，藉此得知三種模型的訓練表現。

在本節中，3.5.1 為 LOKI 訓練和驗證程序，說明 LOKI 模型的建立步驟和操作程序；3.5.2 為簡單貝氏分類器訓練和驗證程序，說明簡單貝氏分類器模型的語料分配和訓練過程；3.5.3 為 LSTM 訓練和驗證程序，說明 LSTM 模型的語料分配和訓練過程。

3.5.1 LOKI 訓練和驗證程序

在處理文本時，自然語言處理領域大多先把單詞轉為某種數值，並基於數學理論來進行後續的計算及推論，但本論文考量華語病句的句型結構，因此，不能只使用數學理論的處理方式，句型結構和語意對於華語病句的偵測及修正有其重要性，故同時採用以語言學為基礎的 LOKI 作為模型架構。

關於三種模型的訓練集，LOKI 與簡單貝氏分類器、LSTM 不同之處在於 LOKI 僅需病句語料即可，在上述的 3.3.1 節中，介紹過 LOKI 的運作原理基於語言學的句法分析，以詞組結構律為核心概念，先由 Articut 斷詞後，而取得每個單詞的詞類，詞類決定單詞在句子中的位置，因此，LOKI 能以一個句子即可辨識相同句型的多個句子。然而，病句的句型眾多，且結構複雜，LOKI 能辨識的句型為已建立的病句句型，因此，本論文基於封閉世界假設(Closed-World Assumption，簡稱 CWA)的概念，有一些相關研究，例如 Reiter (1978)等。若 LOKI 成功辨識的句子為病句，反之，若 LOKI 無法辨識，

則視為非病句，但簡單貝氏分類器和 LSTM 則需建立病句和非病句兩種語料。

本論文把已整理的 340 個病句建立成 LOKI 專案，並把病句進行錯誤分類，錯誤主類別為語法、語意、詞彙，每個病句代表一種錯誤句型，故共有 340 個錯誤句型作為錯誤次類別，有關完整的 LOKI 處理程序，實驗流程圖如圖 4。

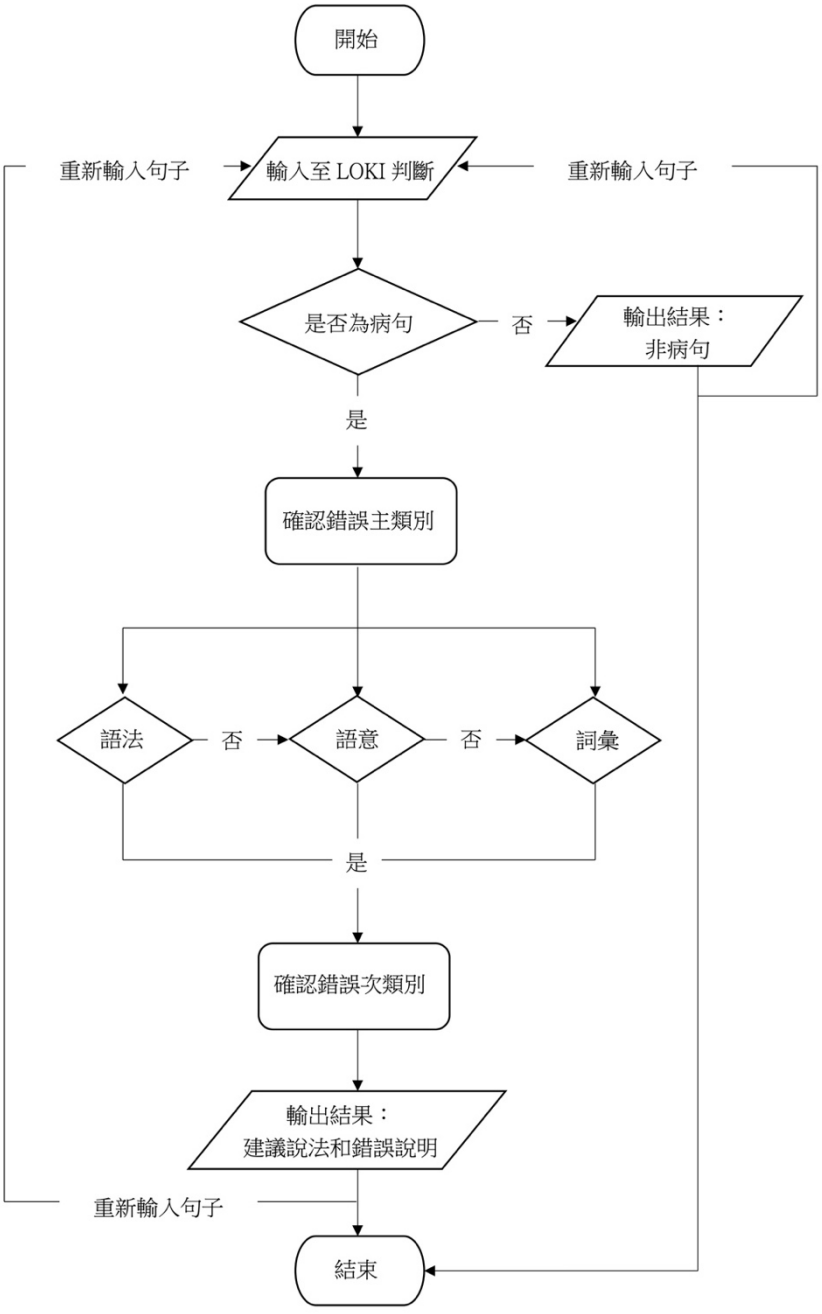


圖 4. LOKI 實驗流程圖

關於語料分配，基於本論文的假設，LOKI 成功辨識的句子視為病句，反之，LOKI 無法辨識的句子則視為非病句，故 LOKI 在建立語料時，只有病句語料，而簡單貝氏分類器和 LSTM 則建立病句和非病句兩種語料，為了維持實驗標準的一致性，需先找出統計式和深度學習的模型其中 20%語料作為驗證集。

第一種分類為病句判斷，若句子為病句，則標籤為 1，反之，若句子為非病句，則標籤為 0。以統計式和深度學習的模型而言，本論文共採用 540 個句子作為語料庫，需先標註正確標籤，因此，340 個病句標註病句標籤為 1，以及 200 個非病句標註非病句標籤為 0，其中 20%語料為驗證集，因此，共有 108 個句子作為 LOKI 的驗證集。接續把驗證集輸入至已訓練的 LOKI 進行判斷，如果 LOKI 辨識為病句，則標示為 1；如果 LOKI 無法辨識，將視為非病句，則標示為 0。

依照 LOKI 回傳的判斷結果與原始的正確標籤進行運算，本論文採用 sklearn.metrics 模組的 accuracy_score 函式和 precision_recall_fscore_support 函式計算正確率等成績，因病句判斷屬於二元分類，故在計算精確率、召回率、F1 分數時，需設定平均方式為 binary，其成績作為 LOKI 的評估指標。

第二種分類為錯誤主類別判斷，確認好句子為病句後，需再確定病句屬於哪種錯誤主類別。為了避免第一種分類病句判斷的結果影響後續的模型判斷，因假設病句判斷有誤判的狀況，後續的模型判斷可能也會出錯，故以統計式和深度學習的模型而言，第二種分類錯誤主類別判斷只採用 340 個病句作為語料庫，並如同上述的語料分配，分成 80%語料作為訓練集，而 20%語料作為驗證集。為了維持實驗標準的一致性，同樣需先找出統計式和深度學習的模型其中 20%語料作為驗證集，共有 68 個病句作為 LOKI 的驗證集。

本論文設定各類別的標籤依序為語法 0、語意 1、詞彙 2，完成設定程序後，把驗證集輸入至已訓練的 LOKI 進行判斷，由 LOKI 歸類病句的錯誤主類別，並回傳其標籤，最後把 LOKI 的判斷結果和原始的正確標籤進行運算，本論文採用如同上述的模組和函式，以 sklearn.metrics 模組的 accuracy_score 函式和 precision_recall_fscore_support 函式計算 LOKI 的評估指標，因錯誤主類別判斷屬於多元分類，故在計算精確率、召回率、F1 分數時，可設定平均方式為 micro 或 macro。

第三種分類為錯誤次類別判斷，確認病句屬於哪種錯誤主類別後，需再確定病句屬於哪種錯誤次類別，才能提供華語病句的建議說法和錯誤說明。以統計式和深度學習的

模型而言，第三種分類錯誤次類別判斷只採用 340 個病句作為語料庫，其原因在於為了避免第一種分類病句判斷的結果影響後續的模型判斷。每個病句代表一種錯誤句型，共有 340 種錯誤句型作為錯誤次類別。為了維持實驗標準的一致性，同樣需先找出統計式和深度學習的模型其中 20%語料作為驗證集，共有 68 個病句作為 LOKI 的驗證集。

因全部病句的數量為 340 個，句型編號達三位數，且為了清楚辨識句型的錯誤主類別，本論文設定第一位數為錯誤主類別的編號，語法編號為 0、語意編號為 1、詞彙編號為 2，故以四位數作為標籤，例如第一句語法病句的標籤為 0001，第一句語意病句的標籤為 1001，第一句詞彙病句的標籤為 2001，以此類推，依序編號作為每種錯誤次類別的標籤，以上為語料處理及訓練程序，主要用訓練集來進行模型訓練。

完成模型訓練後，可用驗證集來評估模型的訓練表現。接續把驗證集輸入至已訓練的 LOKI，由 LOKI 判斷病句屬於哪種錯誤次類別，若有辨識到該句型，則查詢其編號，並回傳編號作為 LOKI 判斷的標籤，最後把 LOKI 的判斷結果和原始的正確標籤進行運算，同樣採用 sklearn.metrics 模組的 accuracy_score 函式和 precision_recall_fscore_support 函式計算 LOKI 的評估指標，因錯誤次類別判斷屬於多元分類，故在計算精確率、召回率、F1 分數時，可設定平均方式為 micro 或 macro。

3.5.2 簡單貝氏分類器訓練和驗證程序

第一種分類為病句判斷，先建立簡單貝氏分類器的語料庫，需病句和非病句兩種語料，這點與 LOKI 不同，LOKI 僅需病句語料，因此，本論文採用 340 個病句和 200 個非病句作為簡單貝氏分類器的語料庫。

把已整理的語料輸入至 Articut 進行斷詞，選用 Articut 作為斷詞系統的原因在於為了維持實驗標準的一致性，所以採用與 LOKI 相同的斷詞系統。接續建立正確標籤，病句為 1，非病句為 0，以便後續用於訓練和驗證的實驗步驟，最後依照簡單貝氏分類器的判斷結果和原始的正確標籤來運算評估指標。

每個病句可視為一個文本，在自然語言處理中，要取得文本特徵，須先把文字轉為數字向量化，機器模型才能辨識。本論文以詞袋模型(Bag of Words，簡稱 BOW)取得文本特徵，我們想像每個文本如同一個袋子，裡面有很多單詞，在詞袋裡不考慮排列順序和語法結構，而是以單詞的出現頻率為主，步驟為先用 sklearn.feature_extraction.text 模

組的 CountVectorizer 函式把文本轉為數字向量化，並提取文本特徵，再用 sklearn.model_selection 模組的 train_test_split 函式進行語料分配，依照自然語言處理的常用做法，分成 80%語料用於訓練，而 20%語料用於驗證，接續選用哪種分類器進行病句判斷，在 sklearn.naive_bayes 模組裡，有分多種分類器，在此採用的是 MultinomialNB 分類器進行模型判斷，最後使用 sklearn.metrics 模組的 accuracy_score 函式和 precision_recall_fscore_support 函式計算評估指標，因病句判斷屬於二元分類，故在計算精確率、召回率、F1 分數時，需設定平均方式為 binary。

訓練步驟依序如下：

1. sklearn.feature_extraction.text 模組的 CountVectorizer 函式：把文本向量化，並提取文本特徵。
2. sklearn.model_selection 模組的 train_test_split 函式：依照語料分配，分成 80%作為訓練集，20%作為驗證集。
3. 把訓練集輸入至簡單貝氏分類器進行訓練。
4. sklearn.naive_bayes 模組的 MultinomialNB 分類器：模型訓練。

驗證步驟依序如下：

1. 把驗證集輸入至已訓練的簡單貝氏分類器。
2. sklearn.naive_bayes 模組的 MultinomialNB 分類器：模型驗證。
3. sklearn.metrics 模組的 accuracy_score 函式和 precision_recall_fscore_support 函式：計算模型的評估指標。

第二種分類為錯誤主類別判斷，為了避免第一種分類病句判斷的結果影響後續的模型判斷，故在錯誤主類別判斷上，只採用 340 個病句作為簡單貝氏分類器的語料庫。由簡單貝氏分類器歸類病句屬於哪種錯誤主類別，錯誤主類別的標籤為語法 0、語意 1、詞彙 2，需先建立好正確標籤，以便後續用於訓練和驗證的實驗步驟，最後依照簡單貝氏分類器的判斷結果和原始的正確標籤來運算評估指標。在語料分配上，同樣是 80%語料用於訓練，而 20%語料用於驗證。因錯誤主類別判斷屬於多元分類，故在計算精確率、召回率、F1 分數時，可設定平均方式為 micro 或 macro，處理程序如同上述。

第三種分類為錯誤次類別判斷，只採用 340 個病句作為簡單貝氏分類器的語料庫，其原因在於為了避免第一種分類病句判斷的結果影響後續的模型判斷。每個病句代表一種錯誤句型，共有 340 種錯誤句型作為錯誤次類別，需先建立好正確標籤，因全部病句的數量為 340 個，句型編號達三位數，且本論文設定第一位數為錯誤主類別的編號，以便清楚辨識句型的錯誤主類別，語法編號為 0、語意編號為 1、詞彙編號為 2，故錯誤次類別以四位數作為標籤，例如第一句語法病句的標籤為 0001，第一句語意病句的標籤為 1001，第一句詞彙病句的標籤為 2001，以此類推，依序編號作為每種錯誤次類別的標籤。

接續由簡單貝氏分類器判斷病句屬於哪種錯誤次類別，最後依照簡單貝氏分類器的判斷結果和原始的正確標籤來運算評估指標。在語料分配上，同樣是 80%語料用於訓練，而 20%語料用於驗證。因錯誤次類別判斷屬於多元分類，故在計算精確率、召回率、F1 分數時，可設定平均方式為 micro 或 macro，處理程序如同上述。

3.5.3 LSTM 訓練和驗證程序

第一種分類為病句判斷，LSTM 的語料庫和簡單貝氏分類器一樣，需建立病句和非病句兩種，因此，同樣採用 340 個病句和 200 個非病句，共有 540 個句子作為 LSTM 的語料庫，並從中分成 80%語料作為訓練集，20%語料作為驗證集。

本論文採用 gensim.models.word2vec 模組的 Word2Vec 函式提取文本特徵，Word to Vector 模型(簡稱 Word2Vec)的功能在於把單詞轉成數字向量化。Word2Vec 是一種詞嵌入(word embedding)的方式，較傳統的方式為使用詞袋來提取文本特徵，其缺點為不考慮排列順序和語法結構，而 Word2Vec 改善此問題，也就是 Word2Vec 考慮鄰近的單詞對語意的影響，在訓練詞向量時，通常同時考慮五個單詞之間的關係。

除此之外，本論文所使用的 Word2Vec 來自國立中正大學 110 學年度第二學期語言學研究所的課程「Python 與自然語言處理(二)」，該詞向量的語料庫來自課堂所搜集的語料，此語料庫共有三千萬字元，使用 CKIP Tagger 進行斷詞，其結果得到 124068 個單詞，並運用連續詞袋(Continuous Bag of Words，簡稱 CBOW)的方式進行訓練。

接續使用 tensorflow.keras.utils 模組的 pad_sequences 函式進行裁剪及填補語料長度，再用 sklearn.model_selection 模組的 train_test_split 函式把語料庫分成訓練集和驗證集，

依照自然語言處理的常用做法，分成 80%語料作為訓練集，而 20%語料作為驗證集。

在神經網路的設定方面，使用 keras.models 模組的 Sequential 函式設定基本條件，例如嵌入維度、世代等細項，再用 keras.layers 模組的 Dense 函式、Dropout 函式、Flatten 函式、LSTM 函式，設定 LSTM、輸出層等閘層，才能進行病句判斷，最後用 tensorflow.keras.metrics 模組的 Precision 函式和 Recall 函式計算評估指標。

訓練及驗證步驟依序如下：

1. gensim.models.word2vec 模組的 Word2Vec 函式：把文本向量化，並提取文本特徵。
2. tensorflow.keras.utils 模組的 pad_sequences 函式：裁剪及填補語料長度。
3. sklearn.model_selection 模組的 train_test_split 函式：依照語料分配，分成 80%作為訓練集，20%作為驗證集。
4. keras.models 模組的 Sequential 函式：設定神經網路的基本條件。
5. keras.layers 模組的 Dense 函式、Dropout 函式、Flatten 函式、LSTM 函式：模型訓練及驗證。
6. tensorflow.keras.metrics 模組的 Precision 函式和 Recall 函式：計算模型的評估指標。

第二種分類為錯誤主類別判斷，為了避免第一種分類病句判斷的結果影響後續的模型判斷，故在錯誤主類別判斷上，只採用 340 個病句作為 LSTM 的語料庫。此部分屬於多元分類，因錯誤主類別的標籤有三種，故需用 tensorflow.keras.utils 模組的 to_categorical 函式把分類標籤轉為獨熱向量，目的在於辨識多種標籤。如同上述的語料分配，80%語料作為訓練集，而 20%語料作為驗證集，處理程序如同上述。

第三種分類為錯誤次類別判斷，只採用 340 個病句作為 LSTM 的語料庫，其原因在於為了避免第一種分類病句判斷的結果影響後續的模型判斷。每個病句代表一種錯誤句型，共有 340 個錯誤句型作為錯誤次類別。此部分屬於多元分類，所以需用 tensorflow.keras.utils 模組的 to_categorical 函式把分類標籤轉為獨熱向量，目的在於辨識多種標籤。語料分配一樣，80%語料作為訓練集，而 20%語料作為驗證集，處理程序如同上述。

3.6 測試階段

在測試階段，本論文另外搜集 100 個句子作為測試集，測試集分為 50 個病句和 50 個非病句，把測試集分別輸入至已訓練的 LOKI、簡單貝氏分類器、LSTM，由此三種模型分別進行病句相關判斷，並藉由評估指標得知三種模型的測試結果。

在本節中，3.6.1 為 LOKI 測試程序，說明 LOKI 的測試過程；3.6.2 為簡單貝氏分類器測試程序，說明簡單貝氏分類器的測試過程；3.6.3 為 LSTM 測試程序，說明 LSTM 的測試過程。

3.6.1 LOKI 測試程序

第一種分類為病句判斷，測試集共有 100 個句子，需先標註正確標籤，其中 50 個病句標註標籤為 1，以及其中 50 個非病句標註標籤為 0，再把測試集輸入至已訓練的 LOKI 進行測試，因 LOKI 能辨識的句型為已建立的病句句型，故本論文假設 LOKI 成功辨識的句子視為病句，反之，LOKI 無法辨識的句子則視為非病句。

本論文的語料標籤分成病句為 1，而非病句為 0。如果 LOKI 辨識為病句，則標示為 1；如果 LOKI 辨識為非病句，則標示為 0。依照 LOKI 回傳的判斷結果與原始的正確標籤進行運算，以取得評估指標。

接著進入第二種分類錯誤主類別判斷。本論文把錯誤主類別的標籤分成語法 0、語意 1、詞彙 2，因病句的句型眾多，且結構複雜，故 LOKI 能辨識的句型為已建立的病句句型，如果 LOKI 無法辨識某句子，其標籤為 3。完成上述的所有設定程序後，為了避免第一種分類病句判斷的結果影響後續的模型判斷，故只採用 50 個病句作為 LOKI 的測試集，把測試集輸入至已訓練的 LOKI 進行測試，由 LOKI 歸類該病句的錯誤主類別，並回傳其標籤，最後把 LOKI 的判斷結果和原始的正確標籤進行運算，以取得評估指標。

最後第三種分類為錯誤次類別判斷，只採用 50 個病句作為 LOKI 的測試集，其原因在於為了避免第一種分類病句判斷的結果影響後續的模型判斷。接續把測試集輸入至已訓練的 LOKI 進行測試，如果 LOKI 成功辨識，再查詢在訓練階段時所設定的句型編號，並回傳編號作為 LOKI 判斷的標籤；如果 LOKI 無法辨識某句子，其標籤為 3，最

後把 LOKI 的判斷結果和原始的正確標籤進行運算，以取得評估指標。

3.6.2 簡單貝氏分類器測試程序

第一種分類為病句判斷，測試集共有 100 個句子，把已整理的測試集以 Articut 進行斷詞，再標註病句 1 和非病句 0 的標籤，需先建立好正確標籤。完成設定程序後，把測試集輸入至已訓練的簡單貝氏分類器進行測試，最後計算評估指標。

第二種分類為錯誤主類別判斷，錯誤主類別的標籤為語法 0、語意 1、詞彙 2，需先建立好正確標籤，為了避免第一種分類病句判斷的結果影響後續的模型判斷，故只採用 50 個病句作為簡單貝氏分類器的測試集。接續把測試集輸入至已訓練的簡單貝氏分類器進行測試，最後計算評估指標。

第三種分類為錯誤次類別判斷，只採用 50 個病句作為簡單貝氏分類器的測試集，其原因在於為了避免第一種分類病句判斷的結果影響後續的模型判斷。錯誤次類別的標籤為在訓練階段時所設定的句型編號，需先建立好正確標籤，再把測試集輸入至已訓練的簡單貝氏分類器進行測試，最後計算評估指標。

3.6.3 LSTM 測試程序

第一種分類為病句判斷，測試集共有 100 個句子，病句和非病句各 50 個，先把測試集以 Articut 進行斷詞，再標註正確標籤，病句為 1，非病句為 0。完成設定程序後，把測試集輸入至已訓練的 LSTM 進行測試，再使用 sklearn.metrics 模組的 accuracy_score 函式和 precision_recall_fscore_support 函式來計算評估指標，並作為 LSTM 的測試結果。

LSTM 的測試結果未採用 tensorflow.keras.metrics 模組內建的計算函式，此部分有別於訓練階段，LSTM 在訓練過程中，須同時提供訓練集和驗證集，由模型自動計算評估指標，但 LSTM 進行預測時，輸出的結果只會顯示預測的答案，不會自動計算評估指標，此外，LOKI 和簡單貝氏分類器皆採用 sklearn.metrics 模組的計算函式，為了維持實驗標準的一致性，故採用 sklearn.metrics 模組的計算函式來取得 LSTM 的測試結果。

第二種分類為錯誤主類別判斷，為了避免第一種分類病句判斷的結果影響後續的模型判斷，故只採用 50 個病句作為 LSTM 的測試集。錯誤主類別的標籤為語法 0、語意

1、詞彙 2，需先建立好正確標籤。接續把測試集輸入至已訓練的 LSTM 進行測試，再使用 sklearn.metrics 模組的 accuracy_score 函式和 precision_recall_fscore_support 函式來計算評估指標，並作為 LSTM 的測試結果。LSTM 的測試結果未採用 tensorflow.keras.metrics 模組內建的計算函式，其原因如同上述。

第三種分類為錯誤次類別判斷，只採用 50 個病句作為 LSTM 的測試集，其原因在於為了避免第一種分類病句判斷的結果影響後續的模型判斷。錯誤次類別的標籤為在訓練階段時所設定的句型編號，需先建立好正確標籤。接續把測試集輸入至已訓練的 LSTM 進行測試，再使用 sklearn.metrics 模組的 accuracy_score 函式和 precision_recall_fscore_support 函式來計算評估指標，並作為 LSTM 的測試結果。LSTM 的測試結果未採用 tensorflow.keras.metrics 模組內建的計算函式，其原因如同上述。

3.7 結語

本論文把華語病句修正視為分類任務，此部分有別於自然語言處理領域大多視為翻譯任務。本實驗分成兩個階段，依序為訓練階段和測試階段，語料庫共有 640 個句子，其中 540 個句子作為訓練階段的語料庫，以及其中 100 個句子作為測試階段的語料庫。

訓練階段的語料庫分別為 340 個病句和 200 個非病句，並把已整理的語料分成三種錯誤主類別，依序為語法、語意、詞彙，以及每個病句代表一種錯誤句型，共有 340 個錯誤句型作為錯誤次類別，依照自然語言處理的常用做法，分成 80%語料作為訓練集，而 20%語料作為驗證集；測試階段的語料庫是本論文另外搜集的句子，與訓練階段的語料庫有所不同，分別為病句和非病句各 50 個，主要用於評估已訓練的模型之測試結果，以維持實驗標準的完整性。

訓練階段分成訓練模型和驗證模型，完成模型訓練及驗證後，進入測試階段。在每個階段中，模型均進行三種分類，依序為病句判斷、錯誤主類別判斷、錯誤次類別判斷，逐步確認病句的錯誤分類，才能提供華語病句的建議說法和錯誤說明。每個階段以正確率、精確率、召回率、F1 分數作為評估指標，從中比較 LOKI、簡單貝氏分類器、LSTM 此三種模型的訓練表現及測試結果。

整體而言，若只用正確率來評估模型表現是不夠的，正確率的評估重點為在整體模型的判斷結果下，其中模型正確判斷所佔的比例，而不考慮模型誤判的狀況，對於評估

模型表現較無區別力，因此，需再計算精確率和召回率，FP 和 FN 為模型誤判的狀況也需納入考量。精確率的問題在於不考慮 FN，FN 為模型判斷為非病句但實際是病句；召回率則不考慮 FP，FP 為模型判斷為病句但實際是非病句，兩者為互補，此外，F1 分數為精確率和召回率的調和平均數，此成績的意義為精確率和召回率一樣重要。

本論文將在第四章呈現訓練階段和測試階段的模型結果，以及說明 LOKI、簡單貝氏分類器、LSTM 此三種模型在各方面的判斷表現。

第四章 實驗結果及問題討論

4.1 概述

本論文把華語病句修正視為分類任務，並實驗在相同數量的少量語料之模型訓練下，比較 LOKI、簡單貝氏分類器、LSTM 此三種模型的表現。本實驗分為訓練和測試兩個階段，並在每個階段中，這三種模型均進行三種分類，依序為病句判斷、錯誤主類別判斷、錯誤次類別判斷，逐步確認病句的錯誤分類，才能提供華語病句的建議說法和錯誤說明。

本章將報告這三種模型在正確率、精確率、召回率、F1 分數的成績，並解釋這四個數值的意義，以及比較 LOKI、簡單貝氏分類器、LSTM 此三種模型的表現差異，最後討論華語病句相關議題。

本章架構如下：4.2 為訓練表現，說明三種模型的訓練表現；4.3 為測試結果，說明三種模型的測試結果；4.4 為 LOKI 模型誤判的錯誤分析，檢視 LOKI 模型誤判的結果，以及說明其可能因素；4.5 為常見的病句錯誤，從本論文的病句語料中，發現不同的華語學習者可能有相同的偏誤狀況；4.6 為問題與討論，最後提出及討論三個華語病句相關議題，一為華語病句是否有規則，二為不同句型仍屬於同一種錯誤類別，三為錯誤類型的可能來源，並針對這三個議題進行探討；4.7 為結語，總結本章內容。

4.2 訓練表現

在訓練階段，由 LOKI、簡單貝氏分類器、LSTM 此三種模型進行一系列的病句相關判斷，其判斷細分為三種分類，依序為病句、錯誤主類別、錯誤次類別。

第一種分類為病句判斷，由模型判斷句子是否為病句，病句標籤為 1，以及非病句標籤為 0，病句判斷屬於二元分類。本論文共有 640 個句子作為語料庫，其中 540 個句子作為訓練階段的語料庫，以及其中 100 個句子作為測試階段的語料庫。

關於訓練階段的語料庫，依照自然語言處理的常用做法，分成 80%語料作為訓練集，病句判斷共有 432 個句子進行訓練，個別數量為 266 個病句和 166 個非病句，資料

統整如表 5，而 20%語料作為驗證集，共有 108 個句子進行驗證，個別數量為 74 個病句和 34 個非病句，資料統整如表 6。

類別	數量
病句	266
非病句	166
共計	432

表 5. 病句判斷的訓練集

類別	數量
病句	74
非病句	34
共計	108

表 6. 病句判斷的驗證集

LOKI 在訓練過程中，只進行模型判斷，無法自動計算評估指標；簡單貝氏分類器在訓練過程中，也無法自動計算評估指標，因此，LOKI 和簡單貝氏分類器須另外輸入驗證集進行預測，其結果用於評估兩者的訓練表現。LSTM 與上述兩者不同，LSTM 在訓練過程中，須同時提供訓練集和驗證集，由模型自動計算評估指標。

LOKI 和簡單貝氏分類器皆採用 sklearn.metrics 模組的 accuracy_score 函式和 precision_recall_fscore_support 函式進行運算，LOKI 的表現為正確率 0.9722、精確率 0.961、召回率 1、F1 分數 0.9801；簡單貝氏分類器的表現為正確率 0.6852、精確率 0.7703、召回率 0.7703、F1 分數 0.7703。

LSTM 則採用 tensorflow.keras.metrics 模組的 Precision 函式和 Recall 函式進行運算，可得到 LSTM 的正確率、精確率、召回率此三個成績，因 LSTM 沒有提供 F1 分數的計算函式，故自行把 tensorflow.keras.metrics 模組所計算的精確率和召回率代入 F1 分數的公式另行運算，LSTM 的表現為正確率 0.6574、精確率 0.7176、召回率 0.8243、F1 分數 0.7673。

在第一種分類病句判斷上，LOKI 達到九成以上的正確率，簡單貝氏分類器和 LSTM 則是達到六成以上的正確率，相較之下，LOKI 的表現優於簡單貝氏分類器和 LSTM，三種模型的訓練表現相關資料統整如表 7，實驗結果的成績取至小數點後第四位。

任務(一)：是否為病句			
	LOKI	簡單貝氏分類器	LSTM
正確率(Accuracy)	**0.9722**	0.6852	0.6574
精確率(Precision)	**0.961**	0.7703	0.7176
召回率(Recall)	**1**	0.7703	0.8243
F1 分數(F1-score)	**0.9801**	0.7703	0.7673

表 7. 病句判斷的訓練表現

首先，LOKI 的召回率為 1，代表 LOKI 能辨識所有的病句。召回率主要計算在確實為病句的狀況下，其中正確判斷為病句所佔的比例，其著重於 FN 所佔的比例，計算公式為 $\frac{TP}{TP+FN}$，從公式來定義，TP 的意思是模型判斷為病句且確實為病句，FN 的意思是模型判斷為非病句但實際為病句，因分子(TP)和分母(TP 加 FN)的數值都需為 1，才可能得到召回率為 1 的結果，故 FN 須為 0，換言之，LOKI 能辨識所有的病句，且沒有誤判成非病句的狀況。

LOKI 的精確率為 0.961，代表有些非病句被誤判成病句，精確率主要計算在模型判斷為病句的狀況下，其中正確判斷為病句所佔的比例，其著重於 FP 所佔的比例，同樣可從公式來定義，精確率的計算公式為 $\frac{TP}{TP+FP}$，因分子(TP)和分母(TP 加 FP)的數值非為 1，故有少數的 FP 個數，FP 的意思是模型判斷為病句但實際為非病句，FP 是一種模型誤判的狀況，也就是 LOKI 有少數非病句被誤判成病句的狀況。

LOKI 的正確率為 0.9722，其原因和精確率的原因相同，計算正確率的公式為 $\frac{TP+TN}{TP+TN+FP+FN}$，分子的 TP 和 TN 是模型正確判斷的狀況，TP 的意思是模型判斷為病句且確實為病句，TN 的意思是模型判斷為非病句且確實為非病句。從兩者的公式而言，正確率的評估重點是在整體模型的判斷結果下，其中正確判斷所佔的比例，而精確率的評估重點則是在模型判斷為病句的狀況下，其中正確判斷為病句所佔的比例，兩者考慮

的面向不同，所以正確率和精確率的成績也有所不同。

第二種分類為錯誤主類別判斷，本論文把錯誤主類別分成三種，分別為語法、語意、詞彙，由模型判斷病句屬於哪種錯誤主類別，並計算各項評估指標，此部分使用 340 個病句作為語料庫，未使用非病句語料的原因在於為了避免第一種分類病句判斷的結果影響後續的模型判斷，故只採用病句語料。

本論文把 340 個病句分成 80%語料作為訓練集，錯誤主類別判斷共有 272 個病句進行訓練，個別數量為 101 個語法病句、52 個語意病句、119 個詞彙病句，資料統整如表 8，而 20%語料作為驗證集，共有 68 個病句進行驗證，個別數量為 22 個語法病句、17 個語意病句、29 個詞彙病句，資料統整如表 9。

主類別	數量
語法	101
語意	52
詞彙	119
共計	272

表 8. 錯誤主類別判斷的訓練集

主類別	數量
語法	22
語意	17
詞彙	29
共計	68

表 9. 錯誤主類別判斷的驗證集

第一種分類病句判斷屬於二元分類，句子是否為病句，以正確率、精確率、召回率、F1 分數作為評估指標，在第二種分類錯誤主類別判斷上，仍使用此四個成績作為評估指標，但錯誤主類別分為語法、語意、詞彙此三種，錯誤主類別判斷屬於多元分類，因此，計算平均的方式有所不同。

LOKI 和簡單貝氏分類器皆採用 sklearn.metrics 模組的 accuracy_score 函式和 precision_recall_fscore_support 函式進行運算，因錯誤主類別判斷是多元分類，分成語法、語意、詞彙此三種類別，故精確率、召回率、F1 分數有 Micro 和 Macro 之別。LSTM 則採用 tensorflow.keras.metrics 模組的 Precision 函式和 Recall 函式進行運算，其原因在於 LSTM 在訓練過程中，須同時提供訓練集和驗證集，由模型自動計算評估指標，進而得到 LSTM 的正確率、精確率、召回率此三個成績，依照 LSTM 相關文件的說明，計算平均應為 Micro，此外，LSTM 的 F1 分數是把精確率和召回率代入 F1 分數的計算公式另行運算。

首先，說明 Micro 和 Macro 的不同之處，Micro 的計算方式為把每個類別的 TP、TN、FP、FN 的個數分別相加，再依照各項評估指標的公式進行運算，樣本數多的類別對於 Micro 成績的影響較大。以 Micro 成績而言，LOKI 的正確率等各項評估指標皆為 0.9706；簡單貝氏分類器的正確率等各項評估指標皆為 0.4412。這兩種模型的個別評估指標皆相同，由各項評估指標的公式來定義，精確率的計算公式為 $\frac{TP}{TP+FP}$，而召回率的計算公式為 $\frac{TP}{TP+FN}$，如果精確率等於召回率，則 FP 等於 FN，這樣才能讓精確率的分母(TP+FP)和召回率的分母(TP+FN)相等，進而得到兩者相同的數值，然而，正確率也相同，正確率的計算公式為$\frac{TP+TN}{TP+TN+FP+FN}$，延續上述內容，如果 FP 等於 FN，代入正確率的計算公式可變成 $\frac{TP+TN}{TP+TN+2FN}$，且正確率和召回率的數值相同，因此，$\frac{TP+TN}{TP+TN+2FN} = \frac{TP}{TP+FN}$，也就是只有 TN 等於 TP 時，相等的計算公式才能成立，$\frac{2TP}{2TP+2FN} = \frac{TP}{TP+FN}$，簡而言之，在 FP 等於 FN 且 TN 等於 TP 的情況下，才能得到正確率、精確率、召回率此三個數值相等的結果。

接續說明 Macro 的計算方式為獨立計算每個類別的 F1 分數，再取各類別的 F1 分數之平均值，佔少數的類別仍有一定的比例，把各類別視為均等重要。以 Macro 成績而言，LOKI 的表現為正確率 0.9706、精確率 0.9737、召回率 0.9652、F1 分數 0.9691；簡單貝氏分類器的表現為正確率 0.4412、精確率 0.4079、召回率 0.3976、F1 分數 0.382。LSTM 在訓練階段是由模型自動計算評估指標，其表現為正確率 0.3824、精確率 0.5、召回率 0.1324、F1 分數 0.2094。

從錯誤主類別判斷的結果顯示，相較於第一種分類病句判斷，LOKI 的評估指標仍達到九成以上，簡單貝氏分類器和 LSTM 的成績相對不好，可能的原因為當類別越多時，模型的訓練語料也需越多，第一種分類病句判斷屬於二元分類，第二種錯誤主類別

判斷屬於多元分類，本論文把錯誤主類別分成語法、語意、詞彙此三種，並只使用 340 個病句作為語料庫，各類別的病句數量為語法 123 個、語意 69 個、詞彙 148 個，對於統計式和深度學習的模型而言，模型所需的訓練語料量不足。

在第二種分類錯誤主類別判斷上，LOKI 的正確率達到九成以上，簡單貝氏分類器的正確率達到四成以上，LSTM 的正確率達到三成以上，相較之下，LOKI 的表現優於簡單貝氏分類器和 LSTM，三種模型的訓練表現相關資料統整如表 10，實驗結果的成績取至小數點後第四位。

任務(二)：病句屬於哪種錯誤主類別					
	LOKI		簡單貝氏分類器		LSTM
	Micro	**Macro**	Micro	Macro	
正確率(Accuracy)	**0.9706**	**0.9706**	0.4412	0.4412	0.3824
精確率(Precision)	**0.9706**	**0.9737**	0.4412	0.4079	0.5
召回率(Recall)	**0.9706**	**0.9652**	0.4412	0.3976	0.1324
F1 分數(F1-score)	**0.9706**	**0.9691**	0.4412	0.382	0.2094

表 10. 錯誤主類別判斷的訓練表現

第三種分類為錯誤次類別判斷，由模型判斷病句屬於哪種錯誤次類別，同樣使用 340 個病句作為語料庫，未使用非病句語料的原因在於為了避免第一種分類病句判斷的結果影響後續的模型判斷，故只採用病句語料。每個病句代表一種錯誤句型，共有 340 種錯誤句型作為錯誤次類別。

本論文把 340 個病句分成 80%語料作為訓練集，錯誤次類別判斷共有 272 個句子進行訓練，而 20%語料作為驗證集，共有 68 個病句進行驗證。評估指標仍採用與上述相同的模組和函式進行運算，LOKI 和簡單貝氏分類器皆用 sklearn.metrics 模組的 accuracy_score 函式和 precision_recall_fscore_support 函式計算評估指標，因錯誤次類別共有 340 種錯誤句型，錯誤次類別判斷也是多元分類，故仍有 Micro 和 Macro 之別，而 LSTM 則用 tensorflow.keras.metrics 模組的 Precision 函式和 Recall 函式進行運算，其原因在於 LSTM 在訓練過程中，須同時提供訓練集和驗證集，由模型自動計算評估指標，

依照 LSTM 相關文件的說明，計算平均應為 Micro，此外，LSTM 的 F1 分數是把精確率和召回率代入 F1 分數的計算公式另行運算。

在 Micro 成績上，LOKI 的評估指標皆為 0.9706；簡單貝氏分類器的評估指標皆為 0。在 Macro 成績上，LOKI 的正確率為 0.9706、精確率等數值皆為 0.9429；簡單貝氏分類器的評估指標皆為 0。LSTM 在錯誤次類別判斷的訓練表現上，評估指標皆為 0。

LOKI 在錯誤主類別和錯誤次類別的判斷上，Micro 的評估指標皆相同，成績同為 0.9706，可能的原因在於 LOKI 在這兩種分類的判斷上，錯誤判斷的句子皆為特定的病句，且佔多數的類別對 Micro 成績的影響較大，故這兩種分類的判斷結果相同。

以 Macro 成績而言，每個病句代表一種錯誤句型，在錯誤次類別判斷上，共有 340 種錯誤次類別，Macro 把各類別視為均等重要，LOKI 的精確率、召回率、F1 分數皆相同，成績同為 0.9429，可能的原因為每個病句代表一種錯誤句型，也就是每種錯誤次類別的數量都只有一個，其所佔的比例相等，LOKI 錯誤判斷的病句為特定的病句，才可能得到相同的結果。

簡單貝氏分類器和 LSTM 在錯誤次類別判斷的評估指標皆為 0，可能的原因在於簡單貝氏分類器以統計原理為基礎，以及 LSTM 基於深度學習神經網路運作，這兩種學習方式需很多相同句型的句子，模型才能辨識和學習句型結構，因此，兩者所需的語料量大，但在本論文中，每個病句代表一種錯誤句型，只使用 340 個病句作為病句語料庫，換言之，每種錯誤次類別的數量僅有一個，所以簡單貝氏分類器和 LSTM 無法從如此少量語料而得到句型結構的訊息。

若訓練模型所需的語料量大，最關鍵的難點在於語料搜集不易，以及語料標註需花費大量人力和時間，三種模型的運作原理不同，LOKI 基於語言學的句法分析，以詞組結構律為核心概念，先分析句型結構，再比對錯誤句型，以達到少量語料即可辨識大量句子之目標。

在第三種分類錯誤次類別判斷上，LOKI 的正確率達到九成以上，簡單貝氏分類器和 LSTM 的正確率則為零，相較之下，LOKI 的表現優於簡單貝氏分類器和 LSTM，三種模型的訓練表現相關資料統整如表 11，實驗結果的成績取至小數點後第四位。

任務(三)：病句屬於哪種錯誤次類別					
	LOKI		簡單貝氏分類器		LSTM
	Micro	**Macro**	Micro	Macro	
正確率(Accuracy)	**0.9706**	**0.9706**	0	0	0
精確率(Precision)	**0.9706**	**0.9429**	0	0	0
召回率(Recall)	**0.9706**	**0.9429**	0	0	0
F1 分數(F1-score)	**0.9706**	**0.9429**	0	0	0

表 11. 錯誤次類別判斷的訓練表現

4.3 測試結果

完成訓練階段，接續進入測試階段。本論文共有 640 個句子作為語料庫，其中 540 個句子作為訓練階段的語料庫，以及其中 100 個句子作為測試階段的語料庫。

關於測試階段的語料庫，本論文另外搜集 100 個句子作為測試集，其中分別為 50 個病句和 50 個非病句。在測試階段中，如同上述，模型均進行三種分類，第一種分類為病句判斷，確認句子是否為病句；第二種分類為錯誤主類別判斷，錯誤主類別分為語法、語意、詞彙，確認病句屬於哪種錯誤主類別；第三種分類為錯誤次類別判斷，每個病句代表一種錯誤句型，共有 340 種錯誤句型作為錯誤次類別，確認病句屬於哪種錯誤次類別。

接續有關測試階段的計算函式，三種模型皆採用 sklearn.metrics 模組的 accuracy_score 函式和 precision_recall_fscore_support 函式進行計算。LSTM 在測試階段需另外用 sklearn.metrics 模組計算評估指標，此部分與 LSTM 在訓練階段有所不同，其原因在於如果用 LSTM 進行預測，輸出的結果只會顯示預測的答案，不會自動計算評估指標，所以需另外使用 sklearn.metrics 模組進行運算，此外，LSTM 的測試結果未採用 tensorflow.keras.metrics 模組內建的計算函式，其原因在於 LOKI 和簡單貝氏分類器皆採用 sklearn.metrics 模組的計算函式，為了維持實驗標準的一致性，故採用 sklearn.metrics 模組的計算函式。

第一種分類病句判斷為二元分類，而第二種分類錯誤主類別判斷和第三種分類錯誤次類別判斷都屬於多元分類，所以計算平均的方式仍有 Micro 和 Macro 之別，三種模型的測試結果同樣以正確率、精確率、召回率、F1 分數作為評估指標，從中比較 LOKI、簡單貝氏分類器、LSTM 的測試結果。

首先是第一種分類病句判斷，共有 100 個句子作為測試集，病句和非病句的數量各有 50 個句子，資料統整如表 12。

類別	數量
病句	50
非病句	50
共計	100

表 12. 病句判斷的測試集

LOKI 的表現為正確率 0.91、精確率 0.9184、召回率 0.9、F1 分數 0.9091；簡單貝氏分類器的表現為正確率 0.7、精確率 0.6786、召回率 0.76、F1 分數 0.717；LSTM 的表現為正確率 0.62、精確率 0.5833、召回率 0.84、F1 分數 0.6885。

正確率主要是觀察在整體模型的判斷結果下，其中正確判斷所佔的比例，以三種模型的正確率而言，LOKI 正確判斷達九成以上，簡單貝氏分類器達七成，LSTM 達六成以上，相較之下，LOKI 的表現優於簡單貝氏分類器和 LSTM。關於第一種分類病句判斷，三種模型的測試結果相關資料統整如表 13，實驗結果的成績取至小數點後第四位。

任務(一)：是否為病句			
	LOKI	簡單貝氏分類器	LSTM
正確率(Accuracy)	**0.91**	0.7	0.62
精確率(Precision)	**0.9184**	0.6786	0.5833
召回率(Recall)	**0.9**	0.76	0.84
F1 分數(F1-score)	**0.9091**	0.717	0.6885

表 13. 病句判斷的測試結果

接續第二種分類錯誤主類別判斷，此部分只使用 50 個病句作為測試集，其原因在於為了避免第一種分類病句判斷的結果影響後續的模型判斷。錯誤主類別分為語法、語意、詞彙，各類別的病句數量依序為 22 個語法病句、5 個語意病句、23 個詞彙病句，資料統整如表 14。

主類別	數量
語法	22
語意	5
詞彙	23
共計	50

表 14. 錯誤主類別判斷的測試集

有關第二種分類錯誤主類別判斷的測試結果，以 Micro 成績而言，LOKI 的正確率等評估指標皆為 0.9；簡單貝氏分類器的正確率等評估指標皆為 0.8；LSTM 的正確率等評估指標皆為 0.76。以 Macro 成績而言，LOKI 的表現為正確率 0.9、精確率 0.75、召回率 0.6942、F1 分數 0.7203；簡單貝氏分類器的表現為正確率 0.8、精確率 0.711、召回率 0.6451、F1 分數 0.6545；LSTM 的表現為正確率 0.76、精確率 0.5173、召回率 0.5613、F1 分數 0.5318。

LOKI 的 Micro 成績皆為 0.9，由各項評估指標的公式來定義，精確率和召回率同為 0.9，精確率的計算公式為 $\frac{TP}{TP+FP}$，而召回率的計算公式為 $\frac{TP}{TP+FN}$，如果精確率等於召回率，則 FP 等於 FN，這樣才能讓精確率的分母(TP+FP)和召回率的分母(TP+FN)相等，進而得到兩者相同的數值，然而，正確率也同為 0.9，正確率的計算公式為$\frac{TP+TN}{TP+TN+FP+FN}$，延續上述內容，如果 FP 等於 FN，代入正確率的計算公式可變成 $\frac{TP+TN}{TP+TN+2FN}$，且正確率和召回率的數值相同，因此，$\frac{TP+TN}{TP+TN+2FN} = \frac{TP}{TP+FN}$，也就是只有 TN 等於 TP 時，相等的計算公式才能成立，$\frac{2TP}{2TP+2FN} = \frac{TP}{TP+FN}$，簡而言之，在 FP 等於 FN 且 TN 等於 TP 的情況下，才能得到正確率、精確率、召回率此三個數值相等的結果。

LOKI 的 Macro 成績大致達到七成，可能的原因為 Macro 的計算方式把各類別視為均等重要，在本論文的測試集中，語法、語意、詞彙此三種錯誤主類別的語料量不

平均，佔多數的錯誤主類別為詞彙。整體而言，三種模型的 Macro 成績相對較不好，可能的原因為錯誤主類別判斷的測試集數量不平均。

整體而言，LOKI 的正確率達到九成，簡單貝氏分類器的正確率達到八成，LSTM 的正確率達到七成以上，相較之下，LOKI 的表現優於簡單貝氏分類器和 LSTM。關於第二種分類錯誤主類別判斷，三種模型的測試結果相關資料統整如表 15，實驗結果的成績取至小數點後第四位。

任務(二)：病句屬於哪種錯誤主類別						
	LOKI		簡單貝氏分類器		LSTM	
	Micro	**Macro**	Micro	Macro	Micro	Macro
正確率(Accuracy)	**0.9**	**0.9**	0.8	0.8	0.76	0.76
精確率(Precision)	**0.9**	**0.75**	0.8	0.711	0.76	0.5173
召回率(Recall)	**0.9**	**0.6942**	0.8	0.6451	0.76	0.5613
F1 分數(F1-score)	**0.9**	**0.7203**	0.8	0.6545	0.76	0.5318

表 15. 錯誤主類別判斷的測試結果

最後為第三種分類錯誤次類別判斷，此部分只使用 50 個病句作為測試集，其原因在於為了避免第一種分類病句判斷的結果影響後續的模型判斷。有關錯誤次類別判斷的測試結果，以 Micro 成績而言，LOKI 的正確率等評估指標皆為 0.9；簡單貝氏分類器的正確率等評估指標皆為 0.72；LSTM 的評估指標皆為 0。以 Macro 成績而言，LOKI 的表現為正確率 0.9，精確率等評估指標皆為 0.8824；簡單貝氏分類器的表現為正確率 0.72，精確率等評估指標皆為 0.5625；LSTM 的評估指標皆為 0。

以 Micro 成績而言，LOKI 的精確率等評估指標皆為 0.9，與第二種分類錯誤主類別判斷的 Micro 成績相同，可能的原因為在錯誤主類別和錯誤次類別的判斷中，兩者錯誤判斷的病句相同，同時查看 LOKI 在這兩種分類錯誤判斷的病句，其原因確實如此。以 Macro 成績而言，LOKI 的精確率等評估指標皆為 0.8824，Macro 的計算方式把各類別視為均等重要，在第三種分類錯誤次類別判斷中，每個病句代表一種錯誤句型，換言之，各種錯誤次類別的數量只有一個，錯誤判斷的病句所佔的比例相同，且 LOKI 錯誤判斷

的病句為特定的病句，Macro 成績才可能相同。

整體而言，LOKI 的正確率達到九成，簡單貝氏分類器的正確率達到七成以上，LSTM 的正確率則為零，相較之下，LOKI 的表現優於簡單貝氏分類器和 LSTM。關於第三種分類錯誤次類別判斷，三種模型的測試結果相關資料統整如表 16，實驗結果的成績取至小數點後第四位。

任務(三)：病句屬於哪種錯誤次類別						
	LOKI		簡單貝氏分類器		LSTM	
	Micro	**Macro**	Micro	Macro	Micro	Macro
正確率(Accuracy)	**0.9**	**0.9**	0.72	0.72	0	0
精確率(Precision)	**0.9**	**0.8824**	0.72	0.5625	0	0
召回率(Recall)	**0.9**	**0.8824**	0.72	0.5625	0	0
F1 分數(F1-score)	**0.9**	**0.8824**	0.72	0.5625	0	0

表 16. 錯誤次類別判斷的測試結果

4.4 LOKI 模型誤判的錯誤分析

LOKI 基於語言學的句法分析為運作原理，以詞組結構律為核心概念，其強大的特色在於能以一個句子即可辨識相同句型的多個句子。因華語病句的句型眾多，且結構複雜，故不能只使用數學理論的處理方式，而 LOKI 的運作方式符合本論文的考量，但在實務操作上仍有一些限制。

LOKI 主要是根據詞類進行辨識，相同詞類的單詞可出現在句子中的相同位置，所以在斷詞及詞類標註後，就能用相同詞類的單詞來取代原本訓練語料的單詞，但因以詞類作為判斷依據，故有一些選擇性限制(selection restriction)的狀況無法顧及，例如「我昨天吃石頭」為病句，經由 LOKI 分析後，其結果為「ENTITY_pronoun / TIME_day / ACTION_verb / ENTITY_noun」，以相同的句型而言，「我昨天吃餅乾」為非病句，若輸入至 LOKI 進行判斷，LOKI 將會辨識成病句，因「石頭」和「餅乾」的詞類皆為

「ENTITY_noun」，因此，在某些句型上可能會發生模型誤判的狀況。

再者，實際查看本實驗的判斷結果，發現某些句型有相同的誤判狀況，例如在訓練語料中，例句為「我呆在費城」，LOKI 分析此例句的結構，擷取的句型為「pos_pronoun / pos_modifier / pos_location」，其中的 pos_pronoun 和 pos_modifier 根據卓騰語言科技公司的實務分析，兩者屬於可有可無(optional)的成分，只要基於這三個相同詞類的單詞所構成的句子即可對上此句型，故 LOKI 辨識到任何包括地點(pos_location)的句子，就可能被判斷為病句，而形成 FP 的誤判狀況，此外，例句為「這年二月底」，此例句的句型結構較短且沒有動詞，以及包括「時間」元素，如果句子包括「時間」元素，就可能對上此例句，這也是導致誤判的因素之一。

4.5 常見的病句錯誤

本節統整語法、語意、詞彙此三種錯誤主類別的常見病句，因病句數量較多，故每種錯誤主類別皆以兩個病句作為例句。

在本節中，4.5.1 為語法病句，過度類化使用「了」、副詞「也」的位置錯誤、分裂句、缺少「都」此四個常見的錯誤；4.5.2 為語意病句，主要原因是用詞錯誤或語法結構有誤；4.5.3 為詞彙病句，大致可分為近義詞、同音字、音調問題等錯誤。

4.5.1 語法病句

在語法病句中，僅有能理解語意的病句，如果病句同時有語法和語意此兩種錯誤，則會歸類於語意病句。語法病句有四個常見的錯誤，分別為過度類化使用「了」、副詞「也」位置錯誤、分裂句、缺少「都」，依序說明及例句如下。

「了」是華語研究和教學相當關注的議題，許多研究探討「了」的句法結構(Lin 2000, 2003, 2003；Wu 2005；Huang 2009；Wu 2010 等)。Huang (2009)考察完成貌「了」的出現狀況，以符合華語語法的句子為前提下，且不改變其語意，其研究分成三個面向進行觀察，一為必須顯示於句中，二為必須隱藏於句子，三為可自由選擇顯示或隱藏於句中。由於華語的「了」用法複雜，是華語學習者的學習難點之一，以「了」的偏誤狀況而言，本論文的華語學習者皆是英語母語者，英語以過去式的形式來表達過去

發生的事情，華語的「了」被當作英語的過去式過度類化使用，如例句(1)。在例句(1)中，動詞「發現」後加上「了」，但實際用法為動詞「發現」後不需加上「了」，應改成「我發現很多女生的名字用很像的字」以及「春草發現她被騙」。

(1) a. *我發現了很多女生的名字用很像的字

b. *春草發現了她被騙

副詞「也」的位置錯誤，如例句(2)。關於「也」的位置錯誤，依照例句(2)的語意判斷，可能的原因為華語句子是否需對等連接詞這方面的議題，有相關研究(Hole 2004；Yang 2020 等)探討此議題。

(2) a. *也失業率走下坡

b. *也未來可能導致許多問題

分裂句「是...的」句型，雖然分裂句的「是」有時會省略，但從例句(3)來看，華語學習者想用分裂句來表達，但都缺少「是」，應改成「我是不好過的」以及「新冠肺炎的防疫是不太成功的」，語意較完整。

(3) a. *我不好過的

b. *新館肺炎的防疫不太成功的

「都」是華語學習者的常見錯誤之一，有許多相關研究(Cheng 1995；Lin 1998；Cheng 2009 等)。關於「都」的錯誤，如例句(4)，在例句(4)中，華語學習者未在動詞前加上「都」，應改成「每個商店都關門了」以及「很多好玩的地方都關了」。

(4) a. *每個商店關門了

b. *很多好玩的地方關了

4.5.2 語意病句

以語意病句而言，大多的原因在於用詞錯誤或語法結構有誤，而造成無法直接判斷語意，因此，LOKI 會再次提問來確認語意，如果使用者答覆「是」，則輸出結果，提供華語病句的建議說法和錯誤說明。

有關用詞錯誤的狀況，如例句(5)，(5a)其中的「事業」應改為「創業」，(5b)其中的「債務」應改為「貸款」；語法結構有誤，如例句(6)，應改成「他們在晚上看電影」以及「新限制而造成長期效應將會提高社會」。

(5) a. *他們決定事業比就業好

b. *可能有很多學生債務

(6) a. *他們有電影晚上

b. *新限制的長期效應會提高社會

4.5.3 詞彙病句

在詞彙病句中，僅有能理解語意的病句，如果病句同時有詞彙和語意此兩種錯誤，則會歸類於語意病句。以詞彙病句而言，大致可分為三類錯誤，第一類錯誤為用詞錯誤，以近義詞為主，例如「到達/到齊」、「發言/公告」、「解除/緩解」、「經驗/經歷」、「壞/差」、「損害/傷害」等；第二類錯誤為語音相關，同音字例如「在/再」、「呆/待」、「帶/戴」等，音調問題例如「聞/問」、「收/受」、「時/是」等；第三類錯誤為使用簡單且直接的說法，例如「假話/謊」。除此之外，在句子中，若句子有錯字或重複字詞的狀況，皆歸類於詞彙病句。

4.6 問題與討論

本論文把華語病句修正視為分類任務，實驗分成兩個階段，一為訓練階段，二為測試階段，並於每個階段進行三種病句相關分類，依序為病句判斷、錯誤主類別判斷、錯

誤次類別判斷。由模型辨識句子是否為病句，再歸類病句屬於哪種錯誤主類別，接續確認病句屬於哪種錯誤次類別，最後才能提供華語病句的建議說法和錯誤說明。

從 340 個病句語料中，發現不同的華語學習者有相同的偏誤狀況，以及有些不同句型的病句可歸類於相同的錯誤類別，並使用相同的錯誤說明，因此，提出三個華語病句相關議題，依序為華語病句是否有規則、不同句型仍屬於同一種錯誤類別、錯誤類型的可能來源。

在本節中，4.6.1 為華語病句是否有規則，關於病句是獨立存在或可歸納成規則，在自然語言處理領域中，目前大多把華語病句修正視為翻譯任務，間接說明病句是分別獨立的，即使這種方式能達到病句修正的效果，但無法提供病句的錯誤說明；4.6.2 為不同句型仍屬於同一種錯誤類別，自然語言處理相關研究的做法和 LOKI 的處理方式基本上皆使用句型結構進行辨識，從本論文的病句語料中，發現即使是不同句型，仍可歸類於相同錯誤主類別；4.6.3 為錯誤類型的可能來源，若華語學習者的母語背景皆相同，從病句語料中，可初步歸類成幾個錯誤類型，例如近義詞、同音字等錯誤類型。

4.6.1 華語病句是否有規則

在第二章文獻回顧中，有關運用自然語言處理技術於華語病句的研究，可分為病句偵測和病句修正。病句偵測主要由模型辨識句子是否為病句，概念是一種二元分類；而病句修正主要由模型修正華語病句，大多把華語病句修正視為翻譯任務，也就是把病句翻譯成非病句。

上述提到自然語言處理領域大多把華語病句修正視為翻譯任務，本論文認為此方式間接說明病句可能是分別獨立的，而非具有規則性。以本論文的實驗程序而言，先把語料輸入至已訓練的 LOKI，並由 LOKI 進行病句辨識，確認句子為病句，再歸類病句的錯誤主類別，接續比對錯誤句型，確認病句的錯誤次類別。透過一系列的病句相關判斷，逐步確認病句的錯誤分類，最後才能修正華語病句，並輸出結果，提供華語病句的建議說法和錯誤說明。

本論文把華語病句修正視為分類任務，而自然語言處理領域以翻譯概念來進行華語病句修正，以序列對序列為主要的處理方式，雖然這種方式能達到病句修正的效果，且能直接更正其病句錯誤，但無法提供病句的錯誤說明。本論文認為若華語病句系統能提

供錯誤說明，藉此華語學習者能從中初步了解病句的錯誤原因，也能即時更正其錯誤，將有助於實踐電腦輔助語言學習之目標。

4.6.2 不同句型仍屬於同一種錯誤類別

關於自然語言處理與華語病句相關研究所需語料量大，主要原因是自然語言處理領域所採用的模型以統計式或深度學習為主，這兩種模型需大量相同句型的句子，模型才能學習句型結構，而 LOKI 的運作原理基於語言學的句法分析，先經過句法分析的過程，取得每個句子的句型結構，再由 LOKI 比對句型，接著後續的處理程序。

自然語言處理領域的常用做法和 LOKI 的處理方式基本上皆使用句型結構進行辨識，從本論文的病句語料中，發現即使是不同句型，仍可歸類成相同錯誤主類別，且能使用同一種錯誤說明，以下舉例說明。

在例句(7)中，(7a)和(7b)都歸類在語法病句，兩者為不同句型，(7a)是疑問句，(7b)是直述句，但例句(7)皆誤用「了」當成過去式，應改成「你還沒結婚嗎」以及「最大的弟弟搬去大學(附近住了)」。

(7) a. *你還沒結婚了嗎

b. *最大的弟弟搬去了大學

在例句(8)中，(8a)和(8b)都歸類在詞彙病句，兩者的內容主要描述新冠肺炎疫情對美國的影響，其中的「變化」和「變成」為不同詞類的單詞，「變化」為名詞，「變成」為動詞，且兩者的句型也不同，「變化」和「變成」都應改為「改變」較符合語意，也就是「我們的生活和經濟都改變了」以及「馬上改變了美國人的生活」。

(8) a. *我們的生活和經濟都變化了

b. *馬上變成了美國人的生活

在例句(9)中，(9a)和(9b)都歸類在詞彙病句，(9a)的「就餐經歷」應改為「用餐經驗」，(9b)的內容是描述家庭發生很多事情，其中的「經驗」應改為「經歷」，對於美國學生而

言，「經驗」和「經歷」是常見的偏誤現象之一，其原因在於兩者的英語翻譯都是「experience」。

(9) a. *我有幾個難忘的就餐經歷

b. *我的家庭經驗了很多

4.6.3 錯誤類型的可能來源

當學習第二語言時，可能受到第一語言的影響，有許多相關研究關注此議題(Jarvis & Pavlenko 2008；Chung et al. 2019 等)。Jarvis & Pavlenko (2008)探討跨語言影響(Crosslinguistic Influence，簡稱 CLI)或遷移(transfer)的理論發展和相關研究。Chung et al. (2019)回顧第一語言(L1)和第二語言(L2)相關研究，並提到語言遷移是一種複雜的互動過程，涉及認知、語言等多元因素的影響。在整理病句語料時，從中發現不同的華語學習者可能出現相同類別的偏誤狀況，例如動詞後加上「了」當作過去式、副詞「也」的位置錯誤、近義詞、同音字等。有關華語偏誤的文獻，以英語母語者為研究對象，其中有幾個偏誤類型與本論文的病句錯誤相似，例如誤用「了」當成過去式、分裂句等。

董子昀等人(2015)使用「華語學習者語料庫」，該語料庫收錄華語學習者參加「華語文能力測驗(TOCFL)」所寫的作文，其涵蓋 A2、B1、B2、C1 此四個分級的華語程度，他們針對「了」進行偏誤分析，並選用 A2 和 B1 等級的作文，為了解初級和中級的華語學習者對於「了」的學習狀況，且以英語母語者為研究對象，文中提到以句型結構而言，「了」的位置有三種狀況，動詞後、句末、前兩者同時出現。此外，「了」可視為動作完成的標記，以英語為母語的華語學習者可能誤用當成過去式的標記。在本論文的病句裡，也出現多次「了」當作過去式的使用狀況，如例句(10)。

(10) a. *我考慮了退學

b. *新冠肺炎疫情對美國生活有不少了的影響

李家豪(2020)研究美國大學的密集型華語課程，以中級以上的小班課堂為研究對象，小班的教學步驟分為五個階段，分別為課文複習、課文問答、語法操練、情境話題、主

題討論，結合多種教學策略，以提高偏誤回饋的效率，在課文複習階段裡，出現有關「是...的」語法偏誤的現象。「是...的」句型為分裂句，本論文的病句也有此類型的偏誤，如例句(11)，應改成「新冠肺炎的防疫是不太成功的」以及「是沒那麼對抗性的」。

(11) a. *新館肺炎的防疫不太成功的

b. *沒是那麼對抗性的

王莫芳等人(2022)針對英日韓二語學習者使用近義詞偏誤的研究，主要探討華語學習者使用「又」和「再」的偏誤狀況，他們的語料來源為國立臺灣師範大學國語中心的線上水平能力測驗的作文(TOCFL on-line writing corpus)，從中找出以英語、日語、韓語為母語的華語學習者的作文，並基於 James (1998)的五種分類進行偏誤分析，分別為遺漏(omission)、誤選(misselection)、錯序(misordering)、誤加(over-inclusion)、混合(blend)，該研究提到以英語為母語的華語學習者即使達到中級程度，仍是難以辨別「又」、「再」等字詞的用法差異，該研究的內容說明「再」有「然後」的語意，以英語為母語者的華語學習者可能誤解為動作接續的「and then」，以及把副詞「再」當作連詞使用的偏誤狀況，並列舉有關「再」的偏誤狀況，當「再」的英語對應詞為「again」的語意時，華語的「再」位置可放在句前，華語學習者會產生語序錯誤的狀況。

在本論文中，華語學習者也出現「再」的偏誤，雖然偏誤狀況非如同王莫芳等人(2022)的研究，但當「再」為英語「again」的語意時，觀察本論文的病句語料，從中發現華語學習者有誤用同音字的狀況，應使用「再」來表達「動作重複」的語意，卻都誤用「在」，如例句(12)，應改成「我承諾我不會再騙春草」以及「我們可以安排再一次見面」。

(12) a. *我承諾我不會在騙春草

b. *我們可以安排在一次見面

華語學習者基於不同的母語背景而有不同的語言特徵，可透過語料庫來觀察其語言現象。張莉萍(2014)採用 TOCFL 語料庫，並基於中介語對比分析方法(Contrastive Interlanguage Analysis，簡稱 CIA)，針對不同母語背景的華語學習者相關語料庫之間進行比較，著重在於不同母語背景的華語學習者之語言特徵，該研究涵蓋六種不同的母語，

分別為英語、日語、韓語、越語、印尼語、泰語，同時也把語料庫依照這六種母語分成子語料庫(sub-corpora)，該研究的內容說明以英語為母語的華語學習者而言，使用代名詞的情況較多，可能的原因在於英語是主語顯著語言，且英語大多以助動詞或疑問句來表達語氣，並不像華語有句尾語氣詞的用法，該研究的結論提到華語學習者會傾向選用接近母語類型的語言形式，且華語和學習者的母語所對應的詞類可能是影響學習者用詞的因素之一，以及基於文化因素，華語學習者使用華語時，會受到母語遷移的影響。

4.7 結語

以訓練階段的結果而言，三種模型相較之下，LOKI 的表現優於統計式簡單貝氏分類器和深度學習神經網路 LSTM。第一種分類為病句判斷，LOKI 達到九成以上的正確率，此外，LOKI 的召回率為 1，代表 LOKI 能正確辨識全部的病句，而簡單貝氏分類器和 LSTM 皆達到六成以上的正確率。第二種分類為錯誤主類別判斷，LOKI 達到九成以上的正確率;簡單貝氏分類器達到四成以上的正確率;LSTM 達到三成以上的正確率，從結果顯示，相較於第一種分類病句判斷，簡單貝氏分類器和 LSTM 的成績相對不好，因當錯誤類別越多時，模型的訓練語料也需越多，但本論文只使用 340 個病句進行模型訓練，故對於統計式和深度學習的模型所需的語料量不足。第三種分類為錯誤次類別判斷，LOKI 的正確率達到九成以上，而簡單貝氏分類器和 LSTM 的正確率皆為零，統計式和深度學習的模型需大量相同句型句子，模型才能辨識和學習句型結構，因此，兩者所需的語料量大，但本論文的每個病句代表一種錯誤句型，換言之，每種錯誤句型的數量僅有一個，所以簡單貝氏分類器和 LSTM 無法成功學習。

以測試階段的結果而言，三種模型相較之下，LOKI 的表現仍優於簡單貝氏分類器和 LSTM。第一種分類為病句判斷，LOKI 達到九成以上的正確率；簡單貝氏分類器達七成的正確率；LSTM 則達到六成以上的正確率。第二種分類為錯誤主類別判斷，LOKI 達到九成的正確率;簡單貝氏分類器達到八成的正確率;LSTM 達到七成以上的正確率，從測試結果顯示，三種模型在錯誤主類別判斷的 Macro 成績相對較不好，其原因為 Macro 的平均方式把各類別視為均等重要，但本論文的測試集數量不平均。第三種分類為錯誤次類別判斷，LOKI 達到九成的正確率;簡單貝氏分類器達到七成以上的正確率;LSTM 則為零。關於 LOKI 錯誤次類別判斷的測試結果，以 Micro 成績來看，LOKI 的

精確率等評估指標與第二種分類錯誤主類別判斷的數值相同，其原因為在這兩種分類中，LOKI 錯誤判斷的病句皆為特定的病句；以 Macro 成績來看，LOKI 的精確率等評估指標皆相同，其原因為 Macro 的計算方式把各類別視為均等重要，每種錯誤次類別的數量只有一個，也就是每種錯誤次類別所佔的比例相等，且 LOKI 錯誤判斷的病句為特定的病句。

最後談論三個華語病句相關議題，一為華語病句是否有規則，自然語言處理領域大多把華語病句修正視為翻譯任務，但本論文則視為分類任務，模型經過一系列的病句相關判斷，逐步確認病句的錯誤分類，才能提供華語病句的建議說法和錯誤說明；二為不同句型仍屬於同一種錯誤類別，自然語言處理領域主要採用統計式和深度學習的模型，因統計式和深度學習的模型需大量相同句型句子，模型才能辨識和學習句型結構，故需大量語料進行模型訓練，而 LOKI 的運作原理基於語言學的句法分析，先經過句法分析的過程，取得句型結構，總而言之，這三種模型的處理方式基本上皆使用句型結構進行辨識，從本論文的病句語料中，發現即使是不同句型，仍可歸類成相同錯誤主類別，且能使用同一種錯誤說明；三為錯誤類型的可能來源，華語偏誤相關文獻探討以英語為母語的華語學習者可能出現的偏誤現象，本論文的華語學習者同為英語母語者，且文獻所列舉的偏誤與本論文的病句錯誤有些許共同之處，因此，本論文初步推論華語病句可能有規則，且可能的原因在於華語學習者的母語背景相同，此外，從本論文的病句語料中，發現不同的病句句型仍可歸類成相同的錯誤類別。

第五章 結論與未來建議

科技發展相當快速，各領域善用科技來提高工作效率，此概念同時發展於語言教學層面，且越來越多的研究證實，電腦科技確實能輔助語言教學和學習。本論文整合華語教學與自然語言處理技術，應用卓騰語言科技公司的自然語言理解引擎 LOKI 來開發華語病句自動偵測及修正之人工智慧系統，並在相同數量的少量語料下，比較三種模型的表現差異，分別為基於語言學理論的 LOKI、統計式簡單貝氏分類器、深度學習神經網路 LSTM。

第二章為文獻回顧，回顧自然語言處理領域和電腦輔助語言學習領域的文獻，有關自然語言處理技術於華語病句的應用，研究方向可分為病句偵測和病句修正，自然語言處理領域大多把病句偵測視為二元分類，模型僅判斷句子是否為病句，無法完整解釋病句的錯誤為何，且大多把病句修正視為翻譯任務，也就是把病句翻譯成非病句，以序列對序列為主要的處理方式，雖然能達到修正的效果，但無法提供華語病句的錯誤說明，此外，自然語言處理領域的研究主要採用華語能力相關檢定的語料庫作為語料來源，代表研究所需語料量大，但最關鍵的問題在於語料搜集不易，以及語料標註需花費大量人力和時間；電腦輔助語言學習領域大多是科技應用的華語師資培訓課程、善用科技工具融入華語教學課堂、數位華語教材等主題，相較之下，有關華語病句的研究較少見，但熟悉華語學習者的偏誤狀況有助於累積華語教學的實務經驗，因此，華語病句相關研究有其重要性。本論文跨領域結合華語教學與自然語言處理，以華語病句為研究重點，由模型經過一系列的病句相關判斷，最後提供華語病句的建議說法和錯誤說明。

第三章為語料說明及實驗方法，介紹語料來源、華語學習者的學習背景和程度、語料處理程序等內容，以及說明實驗程序和評估指標。本論文把華語病句修正視為分類任務，此部分有別於自然語言處理領域視為翻譯任務。關於模型架構，本論文選用兩種常見的機器學習模型，分別是統計式簡單貝氏分類器和深度學習神經網路 LSTM，此外，本論文考量句型結構和語意對於華語病句的偵測及修正有其重要性，故同時採用基於語言學理論的 LOKI。在相同數量的少量語料下，參照比較三種模型的表現差異。本實驗分成訓練階段和測試階段，並在每個階段中，模型均進行三種分類，依序為病句判斷、錯誤主類別判斷、錯誤次類別判斷，此處理程序的用意在於先判斷句子是否為病句，再

判斷病句屬於哪種錯誤主類別，接續判斷病句屬於哪種錯誤次類別，逐步確認病句的錯誤分類，最後才能提供華語病句的建議說法和錯誤說明。本論文共有 640 個句子作為語料庫，並從中分成訓練集、驗證集、測試集。其中 540 個句子作為訓練階段的語料庫，分別為 340 個病句來自美國某大學華語領航學程的課堂寫作，以及 200 個非病句來自國家教育研究院華語文語料庫與能力基準整合應用系統之華語中介語索引典系統，並依照自然語言處理的常用做法，分成 80%語料進行訓練，共有 432 個句子作為訓練集，而 20%語料進行驗證，共有 108 個句子作為驗證集；其中 100 個句子作為測試階段的語料庫，本論文另外搜集句子作為測試集，病句和非病句各 50 個，語料同樣來自國家教育研究院的華語中介語索引典系統。

第四章為實驗結果及問題討論，在相同數量的少量語料下，參照比較 LOKI、簡單貝氏分類器、LSTM 此三種模型的訓練表現及測試結果。整體而言，LOKI 的表現皆優於簡單貝氏分類器和 LSTM，其原因在於統計式和深度學習的模型需大量相同句型的句子作為訓練語料，模型才能辨識和學習句型結構，但本論文僅使用 340 個病句作為病句的訓練語料，對於常見的機器學習模型而言，訓練語料的數量不足，無法成功學習。除此之外，本論文統整常見的病句錯誤，例如過度類化使用「了」、副詞「也」的位置錯誤、分裂句、缺少「都」、近義詞、同音字、音調問題等偏誤狀況。最後討論三個華語病句相關議題，一為華語病句是否有規則，自然語言處理領域大多把華語病句修正視為翻譯任務，但本論文則視為分類任務，模型經過一系列的病句相關判斷，逐步確認病句的錯誤分類，才能提供華語病句的建議說法和錯誤說明；二為不同句型仍屬於同一種錯誤類別，關於模型的處理方式，基於語言學理論的 LOKI、統計式簡單貝氏分類器、深度學習神經網路 LSTM 此三種模型基本上皆使用句型結構進行辨識。從本論文的病句語料中，發現即使是不同句型，仍可歸類成相同錯誤主類別，且能使用同一種錯誤說明；三為錯誤類型的可能來源，當學習第二語言時，可能會受到第一語言的影響，華語偏誤相關文獻探討以英語為母語的華語學習者可能出現的偏誤現象，本論文的華語學習者同為英語母語者，且文獻所列舉的偏誤狀況與本論文的病句錯誤有些許共同之處，以病句語料作為佐證，本論文初步推論華語病句可能有規則，且可能的原因為華語學習者的母語背景相同，此外，從本論文的病句語料中，發現不同的病句句型仍可歸類成相同的錯誤類別。

本論文認為若華語病句系統能提供錯誤說明，藉此華語學習者能從中初步了解病句

的錯誤原因，也能即時更正其錯誤，將有助於實踐電腦輔助語言學習之目標。整體而言，LOKI 展現語言學理論的實務應用，本論文運用 LOKI 建置華語病句自動偵測及修正之人工智慧系統，其運作原理基於語言學的句法分析，以詞組結構律為核心概念，其強大的特色在於能以一個句子即可辨識相同句型的多個句子。本論文僅以 340 個病句作為 LOKI 的語料庫，實際上，LOKI 能辨識的句子遠超過 340 個病句，然而，病句的句型眾多，且結構複雜，LOKI 能辨識的句子為已建立的句型，但 LOKI 僅需一個病句即可學習其句型結構。

關於未來研究的建議，本論文把華語病句自動偵測及修正視為分類任務，此處理方式是可行的，但當前的做法以建立完整的句子為主，若句子結構越長，代表模型辨識的限制也越多，期許日後有機會繼續搜集病句語料，從中歸納華語病句的語法錯誤，以語法結構為核心概念來建立 LOKI 的句型結構，不僅能增加 LOKI 的病句句型，也能擴大系統規模，將能辨識更多句型的華語病句，進而提升華語病句自動偵測及修正系統的完整性，以提供華語教學相關領域作為參考。

參考文獻

Bai, Jianhua, Li, Cong & Yeh, Wen-Chin. 2019. Integrating Technology in the Teaching of Advanced Chinese. *Journal of Technology and Chinese Language Teaching*, 10(1), 73-90.

Bax, Stephen. 2003. CALL-Past, Present, and Future. *System: An International Journal of Educational Technology and Applied Linguistics*, 31(1), 13-28.

Chang, Yu-Yun, Wang, Po-Ya Angela, Hung, Han-Tang, Khóo, Ka-Sîng & Hsieh, Shu-Kai. 2021. Examine Persuasion Strategies in Chinese on Social Media. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, 108-118. Shanghai: Association for Computational Lingustics.

Chen, Binbin & Zhang, Jingyu. 2022. Pre-Training-Based Grammatical Error Correction Model for the Written Language of Chinese Hearing Impaired Students. *IEEE Access*, 10, 35061-35072.

Cheng, Lisa. 1995. On Dou-Quantification. *Journal of East Asian Linguistics*, 4, 197-234.

Cheng, Lisa. 2009. On Every Type of Quantification Expression in Chinese. In Rathert, M. & Giannakidou, A. (eds.), *Quantification, Definiteness, and Nominalization*, 53-75. Oxford University Press.

Cheng, Yong & Duan, Mofan. 2020. Chinese Grammatical Error Detection Based on BERT Model. In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, 108-113. Suzhou: Association for Computational Linguistics.

Chomsky, Noam. 1957. Syntactic Structures. Berlin: de Gruyter.

Chung, Sheila Cira, Chen, Xi & Geva, Esther. 2019. Deconstructing and Reconstructing Cross-language Transfer in Bilingual Reading Development: An Interactive Framework. *Journal of Neurolinguistics*, 50, 149-161.

Harris, Zellig S. 1945. Discontinuous Morphemes. *Language*, 21, 121-127.

He, Tianxiong, Li, Peifeng & Zhu, Qiaoming. 2018. Identifying Chinese Event Factuality with Convolutional Neural Networks. In Wu, Y., Hong, JF. & Su, Q. (eds.) *Chinese Lexical*

Semantics. CLSW 2017. Lecture Notes in Artificial Intelligence, 10709, 284-292. Berlin: Springer.

Ho, Meng-Ching, Chuang, Ching-Yun, Hsu, Yi-Chun & Chang, Yu-Yun. 2021. Hidden Advertorial Detection on Social Media in Chinese. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, 243-251. Taoyuan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Hochreiter, Sepp & Schmidhuber, Jürgen. 1997. Long Short-Term Memory. *Neural Computation*, 9(8), 1735-1780.

Hole, Daniel P. 2004. Focus Background Marking in Mandarin Chinese: System and Theory behind cái, jiù, dōu and yě. London: Routledge.

Hsin, Shih-Chang, Hsieh, Chia-Ling & Chang-Blust, Laura. 2017. Preservice Teacher Training for Online Chinese Teaching: A Case of Distance Courses for High School Learners. *Journal of Technology and Chinese Language Teaching*, 8(1), 86-103.

Huang, Yu-Ting. 2009. On the Environments for the Obligatoriness, Forbiddance and Optionality of Perfective Le. National Chiayi University. (Master's thesis.)

James, Carl. 1998. Errors in Language and Use: Exploring Error Analysis. New York: Longman.

Jarvis, Scott & Pavlenko, Aneta. 2008. Crosslinguistic Influence in Language and Cognition. New York: Routledge.

Koehler, Matthew J. & Mishra, Punya. 2005. Teachers Learning Technology by Design. *Journal of Computing in Teacher Education*, 21(3), 94-102.

Kuang, Hailan, Wu, Kewen, Ma, Xiaolin & Liu, Xinhua. 2022. A Chinese Grammatical Error Correction Method Based on Iterative Training and Sequence Tagging. *Applied Sciences*, 12(9), 4364.

Lane, Hobson, Howard, Cole & Hapke, Hannes Max. 2019. Natural Language Processing in Action: Understanding, Analyzing and Generating Text with Python. New York: Manning.

Lee, Lung-Hao, Lin, Bo-Lin, Yu, Liang-Chih & Tseng, Yuen-Hsien. 2017. Chinese Grammatical Error Detection Using a CNN-LSTM Model. In Chen, W. et al. (eds.),

Proceedings of the 25th International Conference on Computers in Education, 919-921. New Zealand: Asia-Pacific Society for Computers in Education.

Lee, Lung-Hao, Hung, Man-Chen, Chen, Chao-Yi, Chen, Rou-An & Tseng, Yuen-Hsien. 2021. Chinese Grammatical Error Detection Using Adversarial ELECTRA Transformers. In Rodrigo, M. M. T. et al. (eds.), *Proceedings of the 29th International Conference on Computers in Education*, 111-113. Thailand: Asia-Pacific Society for Computers in Education.

Leung, K. Ming. 2007. Naive Bayesian Classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering 2007, 123-156.

Li, Si, Zhao, Jianbo, Shi, Guirong, Tan, Yuanpeng, Xu, Huifang, Chen, Guang, Lan, Haibo & Lin, Zhiqing. 2019. Chinese Grammatical Error Correction Based on Convolutional Sequence to Sequence Model. *IEEE Access*, 7, 72905-72913.

Lin, Jo-wang. 1998. Distributivity in Chinese and its Implications. *Natural Language Semantics*, 6(2), 201-243.

Lin, Jo-wang. 2000. On the Temporal Meaning of the Verbal-le in Chinese. *Language and Linguistics*, 1(2), 109-133.

Lin, Jo-wang. 2003. Temporal Reference in Mandarin Chinese. *Journal of East Asian Linguistics*, 12(3), 259-311.

Lin, Jo-wang. 2003. Time in a Language without Tense: The Case of Chinese. *Journal of Semantics*, 23(1), 1-53.

Lin, Chin-Hsi, Liu, Haixia & Hu, Ying. 2017. Technology and the Education of Chinese-language Teachers: Where Are We Now? *Journal of Technology and Chinese Language Teaching*, 8(1), 1-15.

Reiter, Raymond. 1978. On Closed World Data Bases. In Hervé Gallaire & Jack Minker. (eds.), *Logic and Data Bases*, 55-76. Boston: Springer.

Ren, Hongkai, Yang, Liner & Xun, Endong. 2018. A Sequence to Sequence Learning for Chinese Grammatical Error Correction. In M. Zhang et al. (eds.), *Natural Language Processing and Chinese Computing. NLPCC 2018. Lecture Notes in Artificial Intelligence*, 11109, 401-410. Berlin: Springer.

Rish, Irina. 2001. An Empirical Study of the Naïve Bayes Classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 3, 41-46.

Shulman, Lee S. 1986. Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15(2), 4-14.

Sung, Ko-Yin & Cheng, Hsiu-Jen. 2017. Chinese Language Learning and Teaching through Desktop Videoconferencing. *Journal of Technology and Chinese Language Teaching*, 8(2), 1-24.

Tian, Ye. 2020. Error Tolerance of Machine Translation: Findings from Failed Teaching Design. *Journal of Technology and Chinese Language Teaching*, 11(1), 19-35.

Tseng, Miao-fen. 2017. The Development of Skills Required for Online Chinese Language Teaching. *Journal of Technology and Chinese Language Teaching*, 8(1), 36-65.

Valdebenito, Mario & Chen, Yalin. 2019. Technology as Enabler of Learner Autonomy and Authentic Learning in Chinese Language Acquisition: A Case Study in Higher Education. *Journal of Technology and Chinese Language Teaching*, 10(2), 61-81.

Wang, Wen-jet, Chen, Chia-jung, Lee, Chia-ming, Lai, Chien-yu & Lin, Hsin-hung. 2019. Linguistics-Oriented Keyword Interface NLU System [Computer program]. from https://api.droidtown.co

Wang, Wen-jet, Chen, Chia-jung, Lee, Chia-ming, Lai, Chien-yu & Lin, Hsin-hung. 2019. Articut: Chinese Word Segmentation and POS Tagging System [Computer program]. from https://api.droidtown.co

Wang, Hongfei, Kurosawa, Michiki, Katsumata, Satoru & Komachi, Mamoru. 2020. Chinese Grammatical Correction Using BERT-based Pre-trained Model. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 163-168. Suzhou: Association for Computational Linguistics.

Wu, Jiun-Shiung. 2005. The Semantics of the Perfective LE and Its Context-Dependency: an SDRT Approach. *Journal of East Asian Linguistics*, 14(4), 299-366.

Wu, Jiun-Shiung. 2010. Interactions between Aspect and Temporal Relations: A Case Study of the Perfective Le. *Language and Linguistics*, 11(1), 65-98.

Xiang, Yang. 2018. Grammatical Error Identification for Learners of Chinese as a Foreign Language. Uppsala University. (Master's thesis.)

Xu, Jun. 2020. Machine Translation for Editing Compositions in a Chinese Language Class: Task Design and Student Beliefs. *Journal of Technology and Chinese Language Teaching*, 11(1), 1-18.

Yang, Zhaole. 2020. Yě, yě, yě: On the Syntax and Semantics of Mandarin yě. Netherlands: Leiden University. (Doctoral dissertation.)

Yeh, Jui-Feng, Hsu, Tsung-Wei & Yeh, Chan-Kun. 2016. Grammatical Error Detection Based on Machine Learning for Mandarin as Second Language Learning. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, 140-147. Osaka: COLING 2016 Organizing Committee.

Zhang, Jinbin & Wang, Heng. 2019. Multi-task Learning for Chinese Word Usage Errors Detection. In *Proceedings of the 3rd IEEE International Conference on Computational Intelligence and Applications (ICCIA 2018)*, 45-48.

Zhao, Zewei & Wang, Houfeng. 2020. MaskGEC: Improving Neural Grammatical Error Correction via Dynamic Masking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(1), 1226-1233.

陳亮光。2008。利用多媒體融入華語文教學提升教師提問技巧之研究。*中原華語文學報*，2，49-67。

陳姮良。2014。應用無縫與翻轉學習模式在中文教學的融合與統整應用。*科技與中文教學*，5(1)，75-82。

董子昀、陳浩然、楊惠媚。2015。以「華語學習者語料庫」為本的「了」字句偏誤分析。*中文計算語言學期刊*，20(1)，79-95。

洪嘉馡。2021。以語料庫為本建置中文語法數位平台及其輔助華語教學。*華語文教學研究*，18(1)，59-87。

洪嘉馡、邱詩雯、宋曜廷、張道行。2018。應用「中文階層式語法庫」於華語文寫作教學的效果評估。*科技與中文教學*，9(2)，40-60。

李詩敏、林慶隆。2018。再探「高興」類近義詞：基於語料庫工具輔助之辨析研究。*華語文教學研究*，15(1)，45-83。

李家豪。2020。華語密集班課堂的教學階段、偏誤回饋與糾錯策略。*華語學刊*，29，9-23。

林翠雲。2012。從 Web2.0 數位工具的創新應用看華語教學設計之重要：以僑務委員會華文網路種子師資班之應用為例。*數位學習科技期刊*，4(4)，1-24。

林翠雲。2013。華語遠距教學實務與模組化教材設計。*中原華語文學報*，11，1-33。

林惠玲、陳正倉。2018。現代統計學。修訂版。臺北：雙葉書廊。

連育仁。2018。華語教材行動互動系統發展與教師科技接受度研究。*華文世界*，122，32-45。

王萸芳、林雪芳、盧淑美。2022。英日韓二語學習者使用華語近義詞 「又」和「再」之偏誤探究。*華語文教學研究*，19(1)，59-93。

熊玉雯、李慧萱、宋曜廷。2014。基於 ACTFL 之華語文寫作評分規準。*華語文教學研究*，11(4)，111-139。

許德寶。2015。CALL 研究中的問題。*科技與中文教學*，6(2)，1-16。

詹衛東。2012。樹庫在漢語語法輔助教學中的應用初探。*科技與中文教學*，3(2)，16-29。

詹衛東、馬騰、田駿、砂岡和子。2015。漢語述補結構數據庫的構建及其可視化研究。*科技與中文教學*，6(1)，1-15。

張于忻。2010。華語文數位教材之模組設計探討。*中原華語文學報*，5，179-198。

張莉萍。2014。不同母語背景華語學習者的用詞特徵：以語料庫為本的研究。*中文計算語言學期刊*，19(2)，53-72。

張莉萍。2022。從基於用法的理論探討中高級華語教學語法點。*華語文教學研究*，19(2)，33-64。

曾妙芬、高燕、蔡羅一。2019。中文線上課堂有效結合科技工具以強化互動之報告。*科技與中文教學*，10(1)，91-113。

鄭琇仁。2009。多媒體華語教師的現況與師資培育課程關係探討。*中原華語文學報*，3，107-127。

鄭琇仁。2014。線上華語師資培訓與科技教學學科知識養成之研究。*科技與中文教學*，5(2)，1-18。