

國立清華大學

碩士論文

中文語意搭配詞之預測

Prediction of Chinese Meaningful Word Pairs



系別：資訊系統與應用研究所

學號：111065521

研究生：何品翰(Pin-Han Ho)

指導教授：許聞廉 博士(Prof. Wen-Lian Hsu)

中華民國一一三年七月

# 摘要

此篇論文的主題是中文語意搭配詞之預測，而此方法主要應用為簡化法。首先簡化法的目的在於將複雜句簡化成簡單句，將具有修飾詞的句子做縮減以取得核心的部分，藉此取得主詞、動詞、受詞的主要結構(SVO 結構)。另外透過找到詞和詞的修飾關係，能夠收集句子中特定目標詞的修飾語，形成依存的結構。透過上述將複雜句簡化成簡單句的結果，能夠抓到句子的結構及修飾關係，有利於後續語法分析器(parser)做使用。

而此篇之中文語意搭配詞主要是在收集句子中的具有修飾關係的詞彙組合，目前收集的修飾類別有 5 種。在方法部分採用神經網路模型加上知識庫的方法來綜合預測修飾的關係。模型部分採用 BERT 模型為預訓練模型加上分類器，為多類別的預測模型；在知識庫部分則為收集常見的修飾關係詞彙組合，用以預測可能的修飾關係。最後綜合兩者結果以產出最後的預測關係及類別。

在資料集方面，語料來源採用哈工大語料及小學數學語料。此兩份語料在製作及標註方面均為本研究產生。而本研究在語料標註方面，修飾關係部分是基於哈工大的語法分析工具採用自動標註方式，收集有修飾關係的詞彙組合。在非修飾關係部分則採用選取目標詞(名詞和動詞)前後特定範圍內的詞彙和目標詞的組合來做收集。

而實驗結果部分，目前結果表明，結合知識庫和神經網路模型的方法在預測修飾關係方面較單獨使用知識庫或單獨使用神經網路有較佳之準確度。在知識庫中透過詞彙組合的類別篩選，篩選出單一修飾關係類別的詞彙組合，能顯著提升模型的綜合預測效果。另外自動標註方法的準確率均達到 99%以上，也驗證了自動標註的有效性和可靠性。

**關鍵字：**簡化法、中文語意搭配詞、知識庫、自動標註

# Abstract

The topic of this paper is the Prediction of Chinese Meaningful Word Pairs, primarily applying a reduction method. The purpose of the reduction method is to simplify complex sentences into simple ones, reducing sentences with modifiers to obtain the core parts, in order to acquire the main structure of subject, verb, and object (SVO structure). Additionally, this reduction method allows collecting modifiers for specific target words in sentences, forming a dependency structure. Through the results mentioned above of simplifying complex sentences into simple ones, the structure and modifying relationships of sentences can be captured, which is beneficial for subsequent parsing.

In this paper, Chinese meaningful word pairs primarily focus on collecting word pairs with modifying relationships within sentences. Among the collected modifying relationships, five types of modifying categories have been collected. In the methodology section, a combination of neural network models and knowledge bases is used to predict modifying relationships. The model part uses the BERT model as a pre-trained model plus a classifier, forming a multi-class prediction model. The knowledge base part collects common modifier combinations to predict possible modifying relationships. Finally, the results of both are combined to produce the final predicted relationships and categories.

Regarding the dataset, the corpus sources are from the Harbin Institute of Technology (HIT) corpus and elementary school mathematics corpus. Both of these corpora were created and annotated for this research. In terms of corpus annotation, the modifying relationships were automatically annotated based on the HIT's parser, collecting word pairs with modifying relationships. For non-modifying relationships, word pairs were collected by selecting words within a specific range before and after the target words (nouns and verbs).

The experimental results show that the method combining knowledge bases and neural network models outperforms in predicting modifying relationships compared to using knowledge bases or neural networks alone. In the knowledge base, filtering word pairs by category, selecting word pairs with a single modifying relationship category, significantly improves the model's overall prediction performance. Furthermore, the accuracy of the automatic annotation method consistently reaches over 99%, validating its effectiveness and reliability.

**Keyword: Reduction Method, Chinese Meaningful Word Pairs, Knowledge Base, Auto-labeling**

# 誌謝

在這篇論文結束的同時，也代表著兩年的碩士生涯即將告一段落。而在這兩年的碩士生涯中，我從學校課程、實驗室的研究、同學間的交流中成長了許多。讓我從一個懵懵懂懂的大學生，成長為一個具有獨立研究經驗的研究生。在這段期間，我不僅學到了豐富的專業知識，還培養了面對困難時的堅韌和解決問題的能力。

而在這篇碩士論文的完成過程中，有許多人的幫助和支持，使我能夠順利完成這項研究。在此，我要向所有給予我支持和幫助的人表達最誠摯的謝意。首先，我要感謝我的指導教授許聞廉教授，感謝教授最初的知遇之恩，願意收我為實驗室的一份子，對此心中十分的感激。另外也感謝教授在我研究過程中的悉心指導。許聞廉教授不僅在學術上給予了我寶貴的建議和指導，還給予了我完成研究的信心，使我能夠克服種種挑戰，順利完成論文。而另外我也要感謝實驗室中的朝鈞，他對我的幫助甚至可以尊稱他為我的師父，感謝他在我研究過程中的指導。他扎實的理論基礎和豐富的實作經驗是我完成這篇論文十分重要的推力，可以說沒有他就沒有如今的結果。此外也感謝他願意不厭其煩的讓我詢問問題，不管是研究方面或是人生方面的問題，他總是能迅速的為我解惑，對此幫我省下了不少的時間，也常常讓我能夠突破自己思考的盲點，學到許多分析問題的能力以及正確的思考方式。

其次，我要感謝我的家人，感謝他們在我求學期間的支持和理解。特別是我的父母，他們一直以來的鼓勵和支持是我前進的最大動力。無論是面對學業上的壓力還是生活中的困難，他們總是以無限耐心陪伴著我，給予我支持和鼓勵。此外，在生活上也能給予我充足的資源，讓我能夠專心的做研究、學習，而不用擔心到生活的經濟、財務來源。感謝我的家人，感謝你們一直以來的支持和鼓勵，讓我能夠往自己夢想的道路前進。此外，我還要感謝我的同學和朋友們，感謝他

們在我研究過程中的幫助和陪伴。我們一起討論問題、分享心得，這些經歷讓我的研究生生活變得更加豐富和有趣。另外，每當我遇到學術或生活上的困難或瓶頸時，他們總是願意與我一同探討，提供寶貴的建議和不同的視角，幫助我找到解決問題的方法。再來，我要感謝清華大學和資應所的所有老師和工作人員，感謝他們提供的優良學習環境和資源，讓我能夠專心致志地進行研究。

最後，再次感謝所有在我碩士研究過程中給予我幫助和支持的人，這篇論文的完成離不開你們的支持和鼓勵。



# Table of Contents

摘要.....	i
Abstract.....	ii
誌謝.....	iii
<b>第一章 緒論.....</b>	<b>1</b>
1.1 研究動機與目的.....	1
1.2 研究貢獻.....	2
1.3 論文架構.....	2
<b>第二章 相關文獻探討 .....</b>	<b>4</b>
2.1 句子簡化 (Sentence Simplification).....	4
2.2 句子壓縮 (Sentence Compression) .....	5
2.3 關係提取 (Relation Extraction).....	6
2.3.1 關係提取方法介紹與比較.....	6
2.3.2 聯合提取方法(joint extraction methods)介紹 .....	8
2.3.3 關係提取之學習來源.....	9
<b>第三章 方法.....</b>	<b>11</b>
3.1 詞彙組合和語意搭配詞的關係.....	11
3.2 關係分類的目的.....	11
3.3 關係分類的類別.....	12
3.4 關係預測整體流程.....	13
3.5 神經網路模型架構.....	15
3.5.1 預訓練模型 (Pretrained Model).....	16
3.5.2 分類模型 (Classifier).....	17
3.6 知識庫的建立.....	19
3.7 資料集的收集.....	21
3.7.1 收集修飾關係組合方法.....	22
3.7.2 收集非修飾關係組合方法.....	25
3.7.3 關係組合之資料分布.....	28
3.8 神經網路模型之訓練方法.....	30
3.9 神經網路模型參數設定.....	31
<b>第四章 實驗與結果討論 .....</b>	<b>32</b>

4.1 模型效果.....	32
4.2 知識庫篩選比較.....	33
4.3 人工驗證自動標註之正確率.....	37
4.4 錯誤分析.....	38
4.5 整體模型方法和哈工大對應表方法的比較.....	39
<b>第五章 結論與未來展望 .....</b>	<b>41</b>
5.1 結論.....	41
5.2 未來展望.....	41
<b>參考文獻.....</b>	<b>43</b>
<b>附錄.....</b>	<b>47</b>
A. 哈工大依存關係對應修飾關係分類表.....	47
B. 知識庫頻率篩選.....	49
C. 知識庫和 Bigram 的比較.....	52





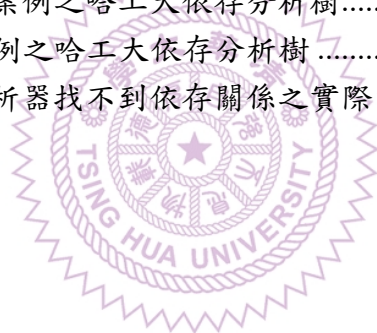
# List of Tables

表格 3.1：關係分類的類別表.....	14
表格 3.2：哈工大語料各類型詞彙組合數量統計.....	21
表格 3.3：小學數學語料各類型詞彙組合數量統計.....	21
表格 3.4：哈工大語料切分資料集數量分布.....	22
表格 3.5：小學數學語料切分資料集數量分布.....	22
表格 3.6：哈工大資料集的 label 分布.....	29
表格 3.7 哈工大資料集的修飾、非修飾關係分布.....	29
表格 3.8：小學數學資料集的 label 分布.....	30
表格 3.9 小學數學資料集的修飾、非修飾關係分布.....	30
表格 3.10：實驗參數設定.....	31
表格 4.1：哈工大語料整體模型效果.....	32
表格 4.2：小學數學語料整體模型效果.....	32
表格 4.3：哈工大語料模型效果比較表.....	33
表格 4.4：小學數學語料模型效果比較表.....	33
表格 4.5：哈工大語料模型效果比較表(知識庫未篩選).....	34
表格 4.6：小學數學語料模型效果比較表(知識庫未篩選).....	34
表格 4.7：哈工大語料模型知識庫類型篩選效果比較表.....	35
表格 4.8：小學數學語料模型知識庫類型篩選效果比較表.....	36
表格 4.9：哈工大語料模型效果比較表(知識庫篩選類型一).....	36
表格 4.10：小學數學語料模型效果比較表(知識庫篩選類型一).....	36
表格 4.11：哈工大語料人工驗證自動標註正確率.....	37
表格 4.12：小學數學語料人工驗證自動標註正確率.....	37
表格 4.13：哈工大語料錯誤分布統計表.....	38
表格 4.14：小學數學語料錯誤分布統計表.....	38
附錄表格 A.1：哈工大依存關係對應修飾關係分類表.....	49
附錄表格 B.1：哈工大語料模型(NN+KB)知識庫頻率篩選效果比較表.....	50
附錄表格 B.2：小學數學語料模型(NN+KB)知識庫頻率篩選效果比較表.....	50
附錄表格 B.3：哈工大語料模型知識庫類型篩選和頻率篩選效果比較表.....	51
附錄表格 B.4：小學數學語料模型知識庫類型篩選和頻率篩選效果比較表.....	52
附錄表格 C.1：bigram 詞性組合對應標註表.....	53
附錄表格 C.2：哈工大語料知識庫和 bigram 比較表.....	53
附錄表格 C.3：小學數學語料知識庫和 bigram 比較表.....	54
附錄表格 C.4：哈工大語料模型(NN+KB) bigram 頻率篩選效果比較表.....	54
附錄表格 C.5：小學數學語料模型(NN+KB) bigram 頻率篩選效果比較表.....	55



# List of Figures

圖片 3.1：修飾關係連線圖.....	12
圖片 3.2：語法分析器(parser)的分析結構示意圖 .....	12
圖片 3.3：關係預測整體流程.....	15
圖片 3.4：神經網路模型架構圖.....	16
圖片 3.5：分類模型架構圖.....	18
圖片 3.6：知識庫整體收集流程.....	20
圖片 3.7：依存分析樹(上)和修飾關係連線圖(下).....	23
圖片 3.8：命名實體處理.....	24
圖片 3.9：收集修飾關係組合整體流程.....	25
圖片 3.10：去除修飾關係組合.....	27
圖片 3.11：去除跨句的組合.....	27
圖片 3.12：去除虛詞的組合.....	27
圖片 3.13：收集非修飾關係組合整體流程.....	28
圖片 4.1：並列結構問題案例之哈工大依存分析樹.....	39
圖片 4.2：pattern 問題案例之哈工大依存分析樹 .....	39
圖片 4.3：哈工大語法分析器找不到依存關係之實際案例.....	40



# 第一章 緒論

## 1.1 研究動機與目的

語言的多樣性為理解和分析帶來了挑戰。自然語言的複雜性不僅來自於其固有的結構和規則，更來自於使用者在不同場合、使用習慣下的表達方式。這樣多變性增加了閱讀者準確把握文本內容的難度。

在當今資訊高度發達的時代，每天產生的文字數據量呈指數級增長。面對如此巨量的資訊，即使是先進的文獻探勘系統也難以全面處理，造成系統負擔。這不僅影響了資訊檢索的效率，也限制了我們從大數據中獲取有價值見解的能力。

然而，通過運用適當的演算法，我們可以將複雜的語言結構簡化為核心語義概念，不僅能夠保留文本的關鍵信息，還能大幅減少需要處理的數據量。這種基於語義的簡化處理方法帶來了多方面的好處：首先，它顯著提高了系統的運行效率，使得同樣的硬體條件能夠處理更多的文本數據。其次，它改善了資訊檢索的準確性，因為系統現在能夠更好地理解文本的實質內容，而不是僅僅依賴表面的詞語匹配的方式。最後，使得研究人員能夠從大量文獻中快速提取有價值的信息。

自然語言的結構複雜性源於其豐富的詞彙和多層次的語法關係。而在簡化法(Reduction method)[1][2]中提到，一個簡單句通常由主語、謂語和賓語等基本成分構成，而這些核心元素往往被各種修飾語、修飾子句和補語所環繞，進而形成了錯綜複雜的語義架構。

語言簡化的核心理念在於保留句子的基本含義，同時精簡其表達形式。具體而言，簡化方法在於識別句子中的核心詞彙及其相關修飾成分。首先修飾關係的部分，可以透過收集詞彙的修飾語集合，找出句子中的修飾成分，進而透過此修飾關係將複雜句反推為原來的簡單句，得出句子的主要結構，即由主語(Subject, S)、謂語(Verb, V)、賓語(Object, O)所組成的結構。而此一過程稱為「簡化」

(reduction)，通過此過程可以將修飾語的語義資訊收納整合到它們所描述的主體詞中。

而這種簡化過程可以通過構建依存分析樹來實現，該樹結構清晰地展示了句子中各詞彙之間的層級關係。從樹的葉節點開始，逐步遞迴地向上合併，將下級詞彙融入上級詞彙中，最終可以達到簡化的目的。此外，這種方法不僅適用於靜態文本處理，還可以擴展到動態的語言交互場景，如語音助理、智能客服系統。

值得注意的是，這種簡化技術具有相當廣泛的適用性，可以應用於不同類型的文本和多種語言（如中文和英文）。不僅能提高文本的可讀性，還能顯著提升自然語言處理系統的效率。在實際應用中，這種方法還可以應用在文本的語言生成。總結來說，自然語言簡化法技術通過提取語言結構，在保留核心語義的同時大幅降低了語言的複雜度，還能夠有助於提高文本處理的效率和準確性。

## 1.2 研究貢獻

本研究主要貢獻可歸納為以下幾個方面：

1. 提供句子中收集具有修飾關係的搭配詞給語法分析器(parser)做使用，以找出句中的修飾關係、輔助其形成依存分析的結構
2. 製作詞彙組合關係預測之資料集
3. 提出詞彙關係預測之神經網路架構、知識庫的預測方式，最後結合知識庫、神經網路進行綜合預測之架構。

## 1.3 論文架構

本論文共分為五個章節，第一章節說明研究動機與目的、貢獻和論文架構；第二章說明句子簡化、句子壓縮一些常見的方法以及關係提取的介紹、方法與其學習來源；第三章則介紹本研究之關係分類整體預測方法及流程，另外詳細介紹

神經網路模型的設計、知識庫的收集以及資料集的製作方式；第四章則提供了關係分類方法的效能評估，包含模型的效果、知識庫的篩選比較、人工驗證自動標註之正確率、錯誤案例分析等等；第五章則是總結了此篇研究的貢獻以及說明未來改進的方向。



## 第二章 相關文獻探討

### 2.1 句子簡化 (Sentence Simplification)

文本簡化 (Text Simplification, TS) 是一項旨在改進文章可讀性的技術。它的目標是調整文字的內容和結構,使其更易於理解,同時保持原文的核心意思。這種技術對於一些特殊群體尤為有用,比如非母語者、語言障礙患者或閱讀困難者[3]。不僅如此,簡化後的文本還能優化其他自然語言處理任務的表現,如語法分析(parsing)、摘要生成(summarization)、資訊擷取 (information extraction) 等。

而在文本簡化領域,目前研究主要聚焦於單句層面的簡化技術,稱為句子簡化 (Sentence Simplification, SS)。這種方法有其優勢:更方便於收集和管理相關的語言數據。而句子簡化和文本簡化相同也是一種旨在提高文本可讀性的語言處理技術,但其目標在於修改單一的句子。其簡化方法涵蓋從詞彙替換到句法重構等方式,包括:插入(insert)、刪除(delete)、取代(change)、改變順序(reorder)、句子分割(split)、句子合併(join)等等[3]。在進行這些轉換的過程中,首要目標是降低句子的複雜度,並確保簡化後的句子仍然符合語法規範,同時保留原文的核心意義。

此外,因為句子簡化的策略多樣,不僅包括縮短句子,還可能會增加文本長度,例如增加連接詞、添加短語做解釋、增加更加明確的指代關係等等。另外,簡化過程還包含刪除(delete)的方法,可能導致資訊刪減,存在影響文本核心語意的風險。因此研究在簡化和保留關鍵信息之間需要找到更好的平衡,以提升句子簡化技術的實用性和有效性。

而目前句子簡化方法中通常透過收集原始和簡化後的句子這種平行語料庫來學習簡化轉換。雖然句子簡化技術已經取得了顯著的進步,但現階段的模型仍面臨著一些挑戰。這些系統雖然已經有了顯著的進展,但仍無法完全達到預期的效能標準。要真正實現對最終用戶有實質幫助的自動化句子簡化,還需要進一步



的技術突破和優化。

## 2.2 句子壓縮 (Sentence Compression)

句子壓縮的主要目的為在不影響句子意思及語法架構的條件下，壓縮、減少句子長度。與句子簡化不同，其主要目的並不是提高句子的可讀性。句子壓縮大部分方法著重於刪除不重要的詞彙

而句子壓縮任務主要包含刪除(Deletion)和抽象(Abstraction)兩種策略。其中刪除導向，主要把句子壓縮當作是序列式的標記或是抽取(Extraction)任務，以去除多餘的詞彙，其中[4]的方法將句子壓縮當作是序列標記的任務，使用雙向長短期記憶(Bi-LSTM)的上下文嵌入和條件隨機場(conditional random field, CRF)用於處理序列標記。而抽象導向方法稱作抽象句子壓縮(Abstractive sentence compression)，其方法主要是透過同義轉換(paraphrasing)來進行原始句子的壓縮，方法包括重新排序(Reordering)、替換(Substitution)、插入(Insertion)，這種方法的主要目標仍是精簡內容，不著重於提升可讀性。另外因為其使用的方法，壓縮後使用的詞彙未必會與原句相同。另外許多方法將神經網路中的機器翻譯模型運用在抽象句子壓縮中，但是效果還遠遠未達到人類滿意的程度[4]。另外因為句子的結構可能提供很重要的資訊，例如主句的主語和動詞可能比介系詞短語或關係子句中的動詞更重要[5]，因此過去還有些方法是使用句法來進行句子的壓縮。

此外，句子壓縮不單只侷限於單句，還有包括多句的句子壓縮方法，類似句子融合。而多句壓縮的目的在於給定一組相似或相關的句子，能夠用一個簡短的句子來概括出它最核心的意義。而其方法主要有抽取的方式，從既有的句子當中，選取有顯著差異的句子，避免形成冗餘的摘要。另外[5]則是提出了建構詞圖的方式，將相關句子的所有單詞中構建一個單詞圖並做壓縮，以生成最核心的摘要。而多句壓縮的應用常常用於生成句子的摘要，例如對話摘要生成或是在有限的空間中顯示摘要，如新聞、電子郵件、維基百科。



## 2.3 關係提取 (Relation Extraction)

要將人類對世界的理解注入到人工智慧系統，知識的收集是關鍵。而實體之間的語意關係是知識的一種表示形式。關係提取(relation extraction, RE)能夠自動化從文本中獲取實體之間的語意關係，是收集知識的重要技術，而它也是資訊提取(information extraction)的一個子任務，在自然語言處理中扮演著重要角色。

關係提取目的為從非結構化的文本中識別實體之間的語義關係，並將其組織為結構化的知識表示。這種結構化知識通常採用關係三元組(relation triples)的形式，描述兩個實體之間的關係。而其具體情況通常為在文本確定實體(entity)後，再根據上下文語境，對這些實體之間的關係進行分類。

另外，關係提取對於構建大規模知識庫和知識圖譜(Knowledge Graphs, KGs)相當重要，這些知識庫和知識圖譜可以支援諸如問答系統(question answering)[6]、對話系統(dialog systems)[7]、搜索引擎(search engine)[8]等下游應用。另外由於網路文本的不斷增長，人類知識也在急速增加，因此有效獲取這些知識對於充實知識庫十分重要。

而關係提取技術的發展經歷了多個階段，傳統的關係提取方法主要是基於模式匹配的方法，而當前主流是採用神經網絡模型方式。在早期的關係提取方法中，其主要在於使用統計相關的方法，例如手工建造模式(hand-built pattern methods)；而後期神經網絡模型又可分為半監督(semi-supervised methods)、監督(supervised methods)、無監督(unsupervised methods)和遠端監督(distant supervision methods)等多種類型。相比於傳統方法中較僵化的字符串模板匹配方式，神經網絡模型能夠更好地擷取文本中的語義信息，在處理複雜的關係提取場景提供更好的效果。

### 2.3.1 關係提取方法介紹與比較

早期的關係提取系統主要依賴於字符串模板的匹配，其運作方式為如果在文本中找到匹配，就可以推斷出該文本具有與匹配模式相對應的關係。然而關係提

取的手動模式構建方法常常需要基於單詞、詞性或語義特徵等等特徵來建立規則，因此常常需要領域專家和語言學家的共同合作，是一個跨領域合作的過程。

而在神經網絡模型中最常使用的則是監督式的方法(supervised methods)，這個方法主要聚焦於使用大量人工標注的數據來訓練模型，以對給定句子中的實體對進行預定義關係的分類。然而，真實情況下的關係提取任務面臨諸多挑戰[9]，包括高品質標注數據取得困難、長尾關係的訓練樣本不足、多句子長上下文的關係提取，以及關係類型的數量成長與既有關係覆蓋度等問題。

根據監督式方法所遇到的種種挑戰，又有了遠端監督方法(distant supervision methods)的提出應用於關係提取任務上。遠端監督屬於弱監督(weakly supervised method)的方法的一種[10]，其利用現有知識庫自動標註訓練數據。其核心思想是：知識庫中存在具有某種關係的實體對，在文本中共同出現時，則該文本段落很可能描述了相同的關係。這一方法的優勢在於能夠快速生成大量帶標籤的訓練數據，大幅減少人工標注的需求。然而，它也面臨著潛在的噪音(noise)問題，因為實際情況上並非所有包含特定實體對的句子都一定表達了知識庫中記錄的那種關係。

而至於半監督方法、無監督方法，半監督關係提取方法利用少量高信心的種子元組(seed tuples)和迭代學習的過程，能高效處理大規模未標記語料，但面臨準確率和語義轉移（逐漸偏離原始關係含義）的問題[10]。無監督關係提取方法則是通過分析實體對的上下文相似性，採用由下而上的策略進行實體對聚類和關係標註，其不需要預先定義關係類型。

關係提取領域存在多種方法，但每種方法都面臨著各自的挑戰：手工模式和半監督方法需要列舉關係模式，容易引入人為偏差。監督方法雖然有成熟的 NLP 工具支持，但仍然較耗費人力與時間。無監督方法產生的結果往往過於寬泛，難以定義合適的關係類別，而且對於處理低頻實體對效果欠佳。遠端監督能高效標記數據，但存在準確率不高的問題。另外整體來說，這些方法大多都侷限在特定領域，且存在錯誤傳播(error propagation)的風險。因此克服這些局限性、提高關

係提取方法的泛化能力和準確度，仍然是此領域的重要研究方向。

### 2.3.2 聯合提取方法(joint extraction methods)介紹

針對錯誤傳播(error propagation)的問題，又有相關研究[11]提出了聯合提取方法(joint extraction methods)。其流程有別於傳統管線化(pipeline)的方式(先取得實體對再預測關係類別)，改成是同時取得的方法(模型同時預測實體對及關係類別)。

此處先介紹 pipeline 的方式，pipeline 的方式採用了兩階段的策略，主要可以分成處理實體和關係抽取兩個步驟。首先，命名實體識別 (Named Entity Recognition，簡稱 NER)，這一步的目標是在文本中找出實體(Entity)及它的範圍並將其進行實體的分類，例如地名、人名、組織名等等。這些識別出的實體可以形成的組合，稱之為實體對，可以用於後續關係提取任務。接著，進行關係抽取。在這一步驟中，系統會根據先前找出的實體對，判斷它們之間是否存在某種關係，並將其進行分類。這種方式的重點在於必須先完成實體的識別工作，然後再進行關係的判斷。換句話說，命名實體識別為關係抽取提供了必要的輸入，兩個步驟是依序進行的。

但是 pipeline 的方式有以下缺點[11]：(1)錯誤傳播(error propagation)：在這個過程中，前一個步驟的錯誤會影響到後續的步驟。結果就是，即使關係抽取的算法本身沒有問題，但因為接收到了錯誤的輸入，最終的結果也會受到影響。(2)資訊遺失：在 pipeline 的方法中，資訊無法有效地在不同任務之間傳遞，影響到最後的預測結果。(3)增加計算成本：在關係抽取的階段，模型會嘗試分析所有可能的實體對之間的關係，包括那些沒有關係的實體對。不僅耗時，還會消耗大量的計算資源，降低整體的處理效率。

而聯合學習方法(joint extraction methods)為應對傳統 pipeline 方法的局限性，其使用混合式的神經網絡，將命名實體識別和關係抽取這兩個原本獨立的任務整

合在一起。相較於流水線方法，聯合學習方法具有兩個顯著優勢：(1)錯誤傳播減少：通過同時處理 NER 和 RE 任務，減少了上游任務錯誤對下游任務的影響。(2)資訊共享：允許兩個子任務之間互相分享資訊，提高整體效果。

而在[11]的研究中提到，句子中一般包含許多的實體對，但真正存在關係的實體對數量卻相對較少。基於這一觀察，此研究不同於傳統先識別實體再確定關係的順序，選擇先提取句子中的關係，定位真正重要的實體對，從而在後續的實體識別過程中更有針對性。其一方面，顯著減少了需要處理的數據量，使得整個計算過程更加高效；另一方面，已知的關係訊息也為實體識別提供資訊，大大提升了識別的準確度。不僅優化了計算效率，還提高了整體結果的預測能力，為實體和關係抽取任務提供了一個更有效率的方法。

### 2.3.3 關係提取之學習來源

而在執行關係提取任務的過程中，實體的名稱和其周圍的上下文都是不可或缺的關鍵信息，兩者共同為關係識別提供了重要依據。實體名稱具有多重功能，不僅能夠透過實體的類別，來幫助限定可能的關係類型，還可用於生成實體嵌入(entity embeddings)，進而輔助關係分類。另外，圍繞實體對的上下文內容所蘊含的語義信息，也是判斷實體間關係的關鍵線索。

而根據人類直覺，上下文應是關係提取的主要判斷來源。然而有相關研究[9]指出模型透過實體名稱可以學習到的知識比上下文更多。其研究通過設計三種不同的實驗設定，深入探討了實體名稱和上下文在關係提取任務中的重要性。這些設置包括：1.一般設定（使用完整的實體名稱和文本）、2.遮蔽實體（masked-entity, ME）設定（用特殊符號替換實體名稱）3.僅實體（only-entity, OE）設定（只提供兩個實體的名稱）。其實驗結果發現在 ME 和 OE 設置下，模型的性能都出現了顯著下降，證實了實體名稱和上下文文本對關係提取任務的重要性。另外，在某些情況下，僅使用實體名稱的表現超過了只使用上下文文本的情況，而此結果違



反人類直覺。因此有不少研究[12]開始對關係提取的判斷依據進行研究，其認為當前的關係提取可能存在一些潛在問題，模型可能無意間利用這些淺層特徵(例如：實體名稱)來做出預測，而不是真正理解文本的深層語境。因此，可能高估了模型解讀上下文的能力，其表現可能不如原先預期的結果。因此此研究為了提升關係提取任務的效能，設計了一個新的預訓練框架，採用實體遮蔽和對比學習的方法。這個框架目的在加深模型對文本上下文的理解，同時有效防止模型僅僅依賴於記憶實體或利用實體名稱的表面特徵。

整體來說，有效的關係抽取需要同時利用上下文和實體信息。未來的研究應致力於更好地整合這兩種信息源，以克服現有數據集可能存在的偏差，實現更準確的關係提取。



## 第三章 方法

在此章節中將介紹本研究中文詞彙組合的關係分類方法，其方法主要包含一個中文詞彙組合的關係分類模型，以及知識庫(Knowledge base)的部分，透過結合兩者來做出關係的預測。另外在本章節將說明本研究資料集的收集和製作方式。

首先在此節中會先介紹詞彙組合和語意搭配詞的關係以及此關係分類方法的目的，接著會對關係分類的類別加以說明，並介紹整體的關係預測流程。再來會介紹中文詞彙組合關係分類模型的部分，並對模型的架構和實作細節詳細說明，包含預訓練模型和分類器。在知識庫(Knowledge base)部分，則會介紹收集及篩選的方式。最後介紹本研究收集、製作資料集的方式，並說明本研究的訓練方法。

### 3.1 詞彙組合和語意搭配詞的關係

在此小節將針對詞彙組合和語意搭配詞兩個名詞進行簡單的定義。首先詞彙組合的定義為句子中任意兩個詞彙的組合，其不一定具有修飾關係。而語意搭配詞的顧名思義為能夠在語意上互相搭配的詞彙組合即屬之，此外這種關聯通常是基於長期使用而形成的慣例。舉例來說，我們習慣說「打了一場漂亮的比賽」，而不會說「打了一場美麗的比賽」，儘管「漂亮」和「美麗」在某些語境下可能有相似的含義。

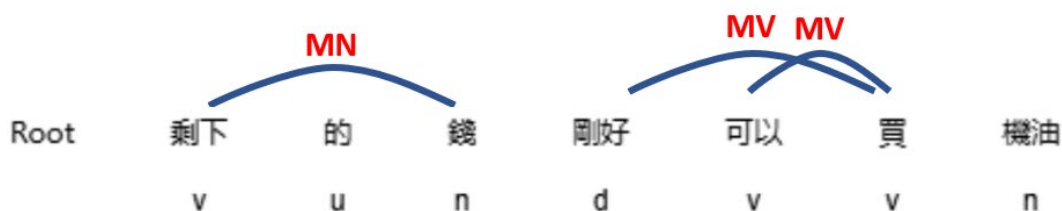
而詞語之間最普遍的搭配關係是修飾關係，例如名詞修飾名詞、形容詞描述名詞等；而對於動詞來說則是地點和時間等副詞或短語。因此此次中文語意搭配詞預測主要目的在於尋找句子中具有修飾關係的詞彙組合。

### 3.2 關係分類的目的

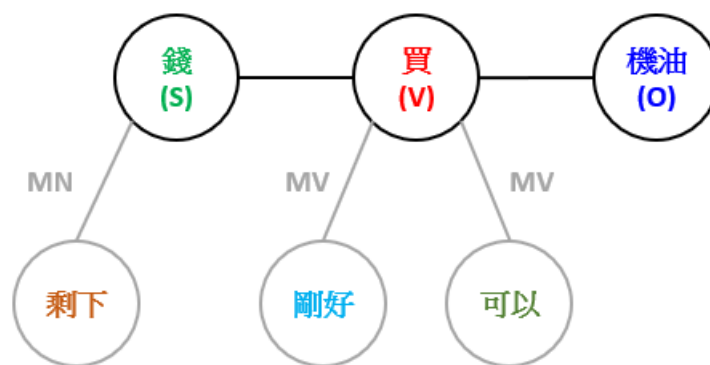
本研究提出之關係分類模型功能在於能夠分類在同一個句子中(不處理跨句的情況)，任兩個詞彙之間的修飾關係。而此目的在於透過找出任一兩兩詞彙之



修飾關係，能夠還原整句的修飾情形，並提供給後續語法分析器(parser)做使用。透過找到整句的修飾關係能夠方便語法分析器(parser)收集目標詞彙的修飾語，產生層級的結構。還能進一步簡化句子，將修飾語向上合併，藉此將複雜句還原成簡單句，呈現主語(Subject, S)、謂語(Verb, V)、賓語(Object, O)的結構，以利整體語法分析器後續的運作。如圖片 3.1、圖片 3.2 所示，透過關係分類模型能夠找出句子中所有兩兩詞彙的修飾關係，藉此還原出整句的修飾關係如圖片 3.1。而透過找出的修飾關係連線，能夠提供語法分析器(parser)分析修飾結構，產生圖片 3.2 的語法分析結構。



圖片 3.1：修飾關係連線圖



圖片 3.2：語法分析器(parser)的分析結構示意圖

### 3.3 關係分類的類別

目前關係分類的類型共有六種，其主要包含 MV、VM、MN、NM、PN、False 共六種(如表格 3.1)，以下分別說明標籤的意義。

首先說明動詞修飾相關的關係標籤，其中 MV 標籤為修飾詞加上動詞的組

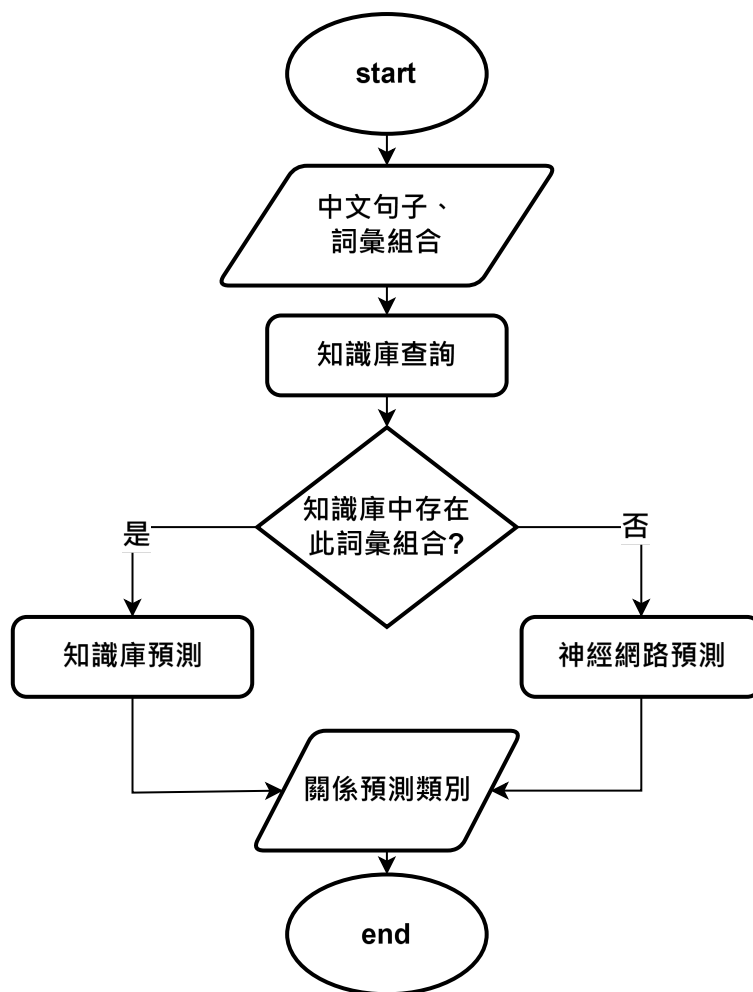
合(modifier + Verb)，主要是修飾詞在動詞前面的情況，例如：(正式，成立)，VM 標籤為動詞加上後方修飾詞的組合(Verb + modifier)，主要是修飾詞在動詞後面的情況，例如：(增長，快)。再來說明和名詞修飾相關的關係標籤，其中 MN 標籤為修飾詞加上名詞的組合(modifier + Noun)，主要是修飾詞在名詞前面的情況，例如：(黃色，香蕉)，NM 標籤為名詞加上後方修飾詞的組合(Noun + modifier)，主要是修飾詞在名詞後面的情況，例如：(石頭，上)。而 PN 標籤為介系詞加上名詞的組合(Preposition + Noun)，主要是介系詞在名詞前面的情況，例如：(在，石頭)。而最後則是 False 標籤，其代表不屬於修飾關係的組合，不屬於上面所述五種的修飾情形即屬於此類別，例如：(啤酒，牛肉)。

### 3.4 關係預測整體流程

本研究之關係預測方法為結合神經網路模型(Neural Network)和知識庫(Knowledge base)之架構，整體關係預測流程如圖片 3.3 所示。首先在輸入層中，輸入為一個中文句子以及當中的一個詞彙組合，接著預測過程中，會先經過知識庫的查詢預測其可能的修飾關係，若知識庫中存在此詞彙組合，則會輸出知識庫預測的結果；若知識庫不存在此詞彙組合，則交由神經網路進行最後的預測，最終輸出層的結果即可得到此預測流程產生的關係預測類別。

關係類別	說明	例子
<b>MV</b>	modifier + Verb	正式 + 成立
<b>VM</b>	Verb + modifier	增長 + 快
<b>MN</b>	modifier + Noun	黃色 + 香蕉
<b>NM</b>	Noun + modifier	石頭 + 上
<b>PN</b>	Preposition + Noun	在 + 石頭
<b>False</b>	非修飾關係	啤酒 + 牛肉

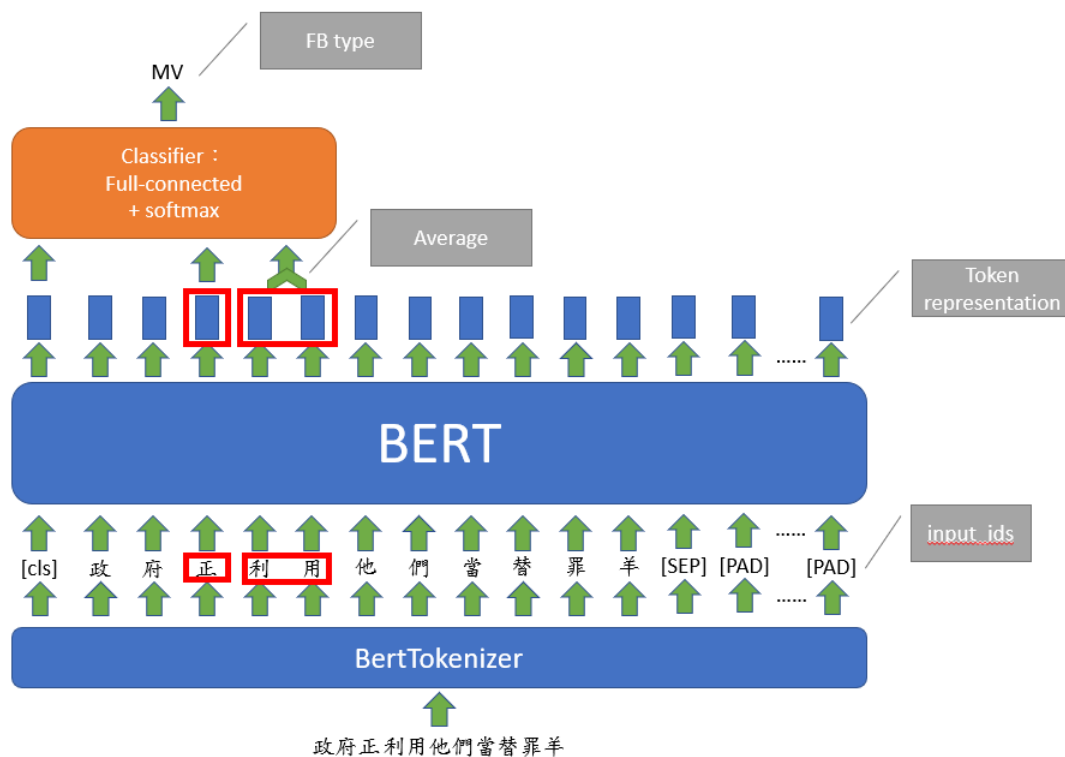
表格 3.1：關係分類的類別表



圖片 3.3：關係預測整體流程

### 3.5 神經網路模型架構

此部分介紹神經網路模型之架構，此模型之架構如圖片 3.4 所示。其架構依序為輸入層、預訓練模型、分類模型、輸出層，其中模型輸入為一個中文句子及當中的一個詞彙組合，輸出為此詞彙組合預測的關係類別，另外預訓練模型、分類模型部分將在以下分別做介紹。



圖片 3.4：神經網路模型架構圖

### 3.5.1 預訓練模型 (Pretrained Model)

本研究之預訓練模型採用哈爾濱工業大學(簡稱哈工大)訊飛聯合實驗室(HFL)開發之中文預訓練模型，此次選用其開發的 RoBERTa-wwm-ext 模型，此模型是一個專門針對中文語言設計的預訓練模型，其核心特點是採用了全詞遮罩 (Whole Word Masking) [13]技術，而其訓練語料為中文維基百科(涵蓋繁體以及簡體)和其他百科、新聞、問答等資料。

此處先解釋何謂全詞遮罩(Whole Word Masking, wwm)，又稱為全詞 Mask 或整詞 Mask，其特別之處在於改變了原來模型預訓練階段中生成訓練樣本的方式。在傳統的 WordPiece 分詞模式下，完整詞彙常被切割成多個子詞單元，在生成訓練資料時，這些子詞單元會被隨機選擇遮蔽 (mask)。然而，全詞遮蔽 (Whole Word Masking) 採用了不同的策略：當一個完整詞彙中的任何 WordPiece 子詞被

選中遮蔽時，該詞的所有相關子詞都會一併被遮蔽[13]。

由於 google 官方發布的 BERT-base 中文模型在處理中文時，採用了以單個字符為基本單位的分割方式。這種方法忽略了中文語言的特殊性，特別是傳統自然語言處理中常用的中文分詞（CWS）技術。相比之下，哈工大開發的預訓練模型採用了更適合中文特性的方法。將全詞遮蔽（Whole Word Masking）技術應用於中文處理中，這種方法能夠更好地維持中文詞彙的完整性和語義。

而此次預訓練模型選用 RoBERTa-wwm-ext 而非其他哈工大開發且同樣採用全詞 Mask 的預訓練模型(如 BERT-wwm, BERT-wwm-ext, RoBERTa-wwm-ext-large...)有以下原因。一方面考量 RoBERTa-wwm-ext 模型在繁體中文閱讀理解任務上有相當不錯之效果(EM 和 F1 平均值分別為 85.2 及 91.7)，另一方面考量到其 base 和 large 版本上參數的差距(base 版本參數數量 110M, large 版本參數數量 330M)，因此最後決定預訓練模型選用 RoBERTa-wwm-ext。而 RoBERTa-wwm-ext 預訓練模型為 base 版本，其結構為 12-layer, 768-hidden, 12-heads, 110M parameters。

而其運作方式是將一個中文句子輸入預訓練模型，而在輸入預訓練模型之前，此句子會先經過預訓練模型的分詞器(Tokenizer)，而分詞器會將句子轉換成一個標記序列(Token sequence)，此序列由一個個的標記(Token)所組成，而此分詞器會在句子的第一個位置加上 CLS 的 Token，在最後句子結尾加上 SEP 的 Token，並做填充(Padding)，再將此標記序列輸入至預訓練模型。而後經過預訓練模型計算即可得到每個 token 位置相對應的上下文詞嵌入(Contextualized Word Embedding)，其包含了此 token 的上下文資訊，此上下文詞嵌入的大小為一個 768 維的向量。

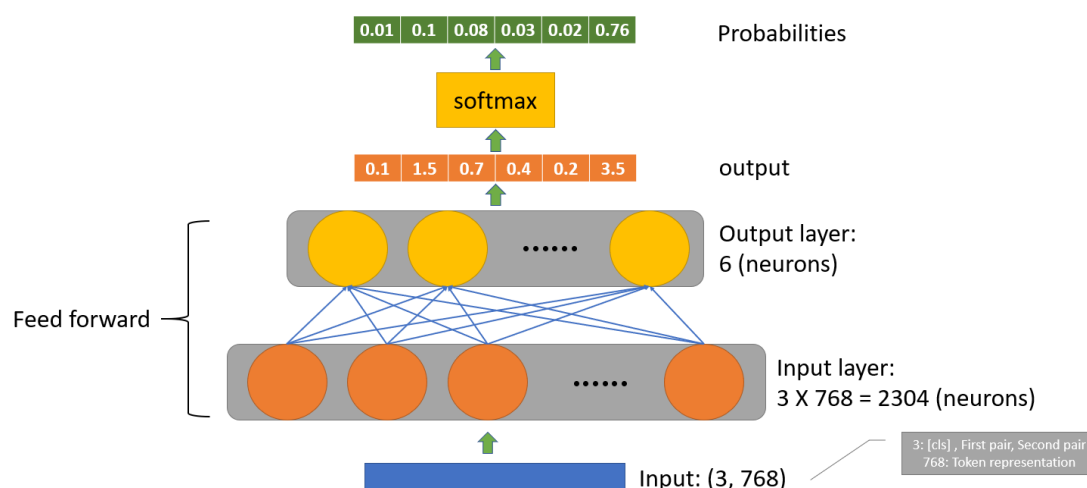
### 3.5.2 分類模型 (Classifier)

接續前面預訓練模型的輸出，可以得到一個上下文詞嵌入的序列，而後續則經由此分類模型得到最後的預測結果，即預測的關係類別。首先，在分類模型輸



入部分，因為此為詞彙組合之關係分類模型，因此需取出詞彙組合位置之上下文詞嵌入，而此詞彙組合固定由兩個詞所組成，因此需要分別取得代表此兩個詞彙的上下文詞嵌入。而其中在每個詞彙部分又可能由一個或多個上下文詞嵌入組成，因此為了得到每個詞彙的上下文詞嵌入並統一每個詞嵌入的大小，因此此處採取相加再取平均的方式，將詞彙中涵蓋的上下文詞嵌入做相加再取平均，即可得到每個詞彙的詞嵌入。而後續將詞彙組合的兩個詞嵌入結合第一個位置的 cls 的上下文詞嵌入，其代表整句話的語義，可以得到一個  $3 \times 768$  的矩陣，此矩陣則為此分類模型的輸入。此  $3 \times 768$  的矩陣主要由三部分的上下文詞嵌入所構成，分別是 cls、詞彙組合的前半部、詞彙組合的後半部，而每一個詞嵌入分別為一個 768 維的向量，因此結合後，即為一個  $3 \times 768$  的矩陣。

而在此分類模型如圖片 3.5 所示，其結構為一層的前饋神經網路 (Feedforward Neural Network)，其輸入層由 2304 個神經元(neuron)所組成，即將輸入矩陣攤平後的大小，而輸出層為 6 個神經元，即關係分類類型的個數。而後續將計算結果經過歸一化指數函數 (softmax function)，即可將分類模型之輸出轉換為關係分類類型的機率分布，最後取最大機率者為此詞彙組合之預測的關係類別。



圖片 3.5：分類模型架構圖

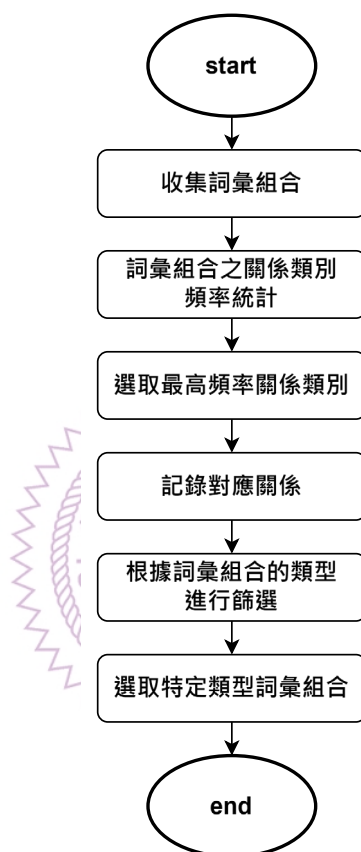
### 3.6 知識庫的建立

知識庫部分則為一個簡單的查詢系統，其預測不考慮句子的上下文，輸入為詞彙組合，而輸出為詞彙組合之預測關係類別。其運作方式為在知識庫中輸入欲查詢的詞彙組合，其根據知識庫中已記錄之詞彙組合查找此詞彙組合及其對應的關係類別並作出預測。例如欲查詢以下中文句子「這條黃色的緞帶」中之(黃色, 緞帶)是否為修飾關係，則方法為在知識庫中輸入此詞彙組合(黃色, 緞帶)，假設知識庫中存在此詞彙組合及其關係類別，則知識庫即可輸出對應的關係類別作預測，在此例中為關係類別 MN (Modifier + Noun)。其描述如下，令輸入為詞彙組合  $P$ ,  $P = \{w_1, w_2\}$ ；輸出為詞彙組合之預測關係類別  $R$ ；KB 為知識庫，其中 KB 包含已記錄的詞彙組合及其對應的關係類別。查詢詞彙組合  $P$  在知識庫中的對應關係類別公式為  $R = KB(P)$ 。

而其收集方式為紀錄資料集中的訓練資料集(train set)部分的詞彙組合，並根據在訓練資料集出現的詞彙組合統計其出現的在各個關係類別的個數，並選取頻率最高的關係類別當作此詞彙組合預測的關係類別，以此來記錄詞彙組合和關係類別的對應關係。而其中若此詞彙組合頻率最高為非修飾關係(False)類別，則知識庫中不進行紀錄，僅收集對應類別是修飾關係(True)的組合，以利後續的預測。

另外此知識庫收集的方法，所收集到的詞彙組合與關係類別的對應關係，每組詞彙組合的預測準確度未必都一致，因此需要透過一些過濾的方式來篩選出真正高品質的詞彙組合和關係類別的對應。因此此處提出詞彙組合類型篩選的方法，根據詞彙組合的類型做出的篩選，以便收集需紀錄的詞彙組合。而此處將詞彙組合分成以下四種類型，單一修飾關係類別、修飾/非修飾關係類別、多修飾關係類別、純非修飾關係類別。以下將分別做說明，首先類型一：單一修飾關係類別是指在訓練資料集中此詞彙組合僅出現一種修飾關係類別則屬之；而類型二：修飾/非修飾關係類別是指此詞彙組合出現修飾和非修飾關係的情況則屬於此類別；另外類型三：多修飾關係類別是指此詞彙組合出現多種修飾關係類別，而沒有出

現過非修飾關係的情況；最後類型四：純非修飾關係類別是指此詞彙組合僅出現過非修飾關係的情況。而此四種類型經過 4.2 小節的實驗(4.2 小節：知識庫篩選方式比較)，得出知識庫詞彙組合選取類型一：單一修飾關係類別搭配神經網路進行預測關係類別的整體效果是最佳的，代表選取此類型的詞彙組合能夠篩選出真正高品質的詞彙組合和關係類別的對應以搭配神經網路進行預測。



圖片 3.6：知識庫整體收集流程

詞彙組合類型	類型一： 單一修飾 關係類別	類型二： 修飾/非修飾 關係類別	類型三： 多修飾 關係類別	類型四： 純非修飾 關係類別
個數	76694	3288	353	102375
比例	41.98%	1.80%	0.19%	56.03%

表格 3.2：哈工大語料各類型詞彙組合數量統計

詞彙組合類型	類型一： 單一修飾 關係類別	類型二： 修飾/非修飾 關係類別	類型三： 多修飾 關係類別	類型四： 純非修飾 關係類別
個數	8905	477	12	22215
比例	28.17%	1.51%	0.04%	70.28%

表格 3.3：小學數學語料各類型詞彙組合數量統計

### 3.7 資料集的收集

此次研究使用之資料集均為本研究製作產生，而此次資料集的語料來源有二：分別為哈工大的語料、小學數學語料。首先，哈工大語料部分為哈工大語法分析器(parser)所使用之原始語料；而小學數學部分則為 IASL(Intelligent Agent Systems Lab)實驗室所提供。而此兩個語料之資料集部分切分之數量及比例如表格 3.4、表格 3.5 所示，其中哈工大語料按照原始語料切分比例分成 train、validation、test 資料集，而小學數學語料按照 8：1：1 比例切分資料集。而接下來幾個小節則是針對資料的標註方式分別做說明。

	<b>Train</b>	<b>Validation</b>	<b>Test</b>
<b>數量</b>	82270	10234	10518
<b>比例</b>	79.86%	9.93%	10.21%

表格 3.4：哈工大語料切分資料集數量分布

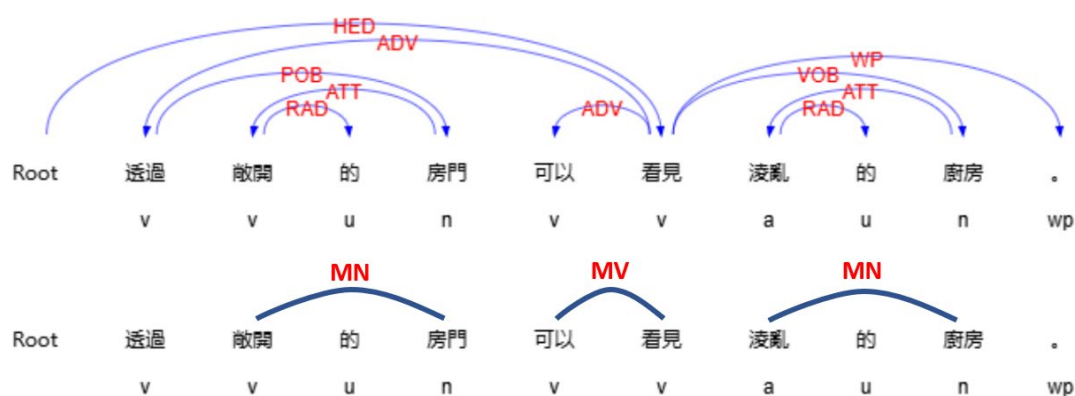
	<b>Train</b>	<b>Validation</b>	<b>Test</b>
<b>數量</b>	266098	5530	10970
<b>比例</b>	94.16%	1.95%	3.88%

表格 3.5：小學數學語料切分資料集數量分布

### 3.7.1 收集修飾關係組合方法

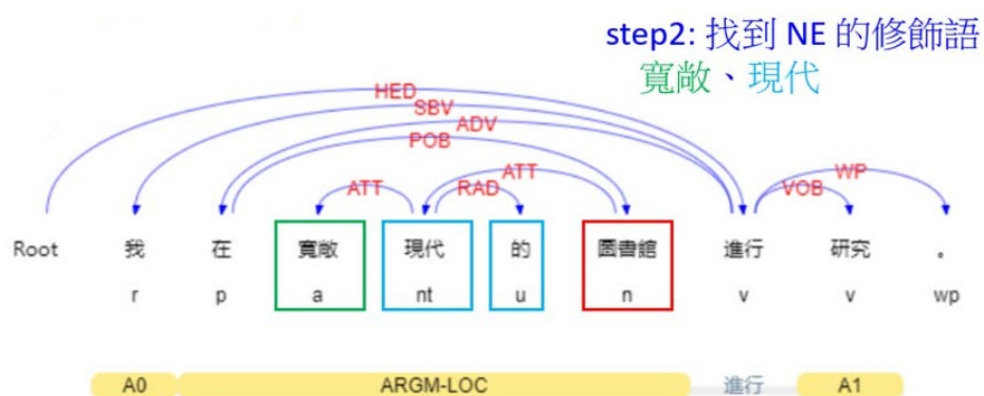
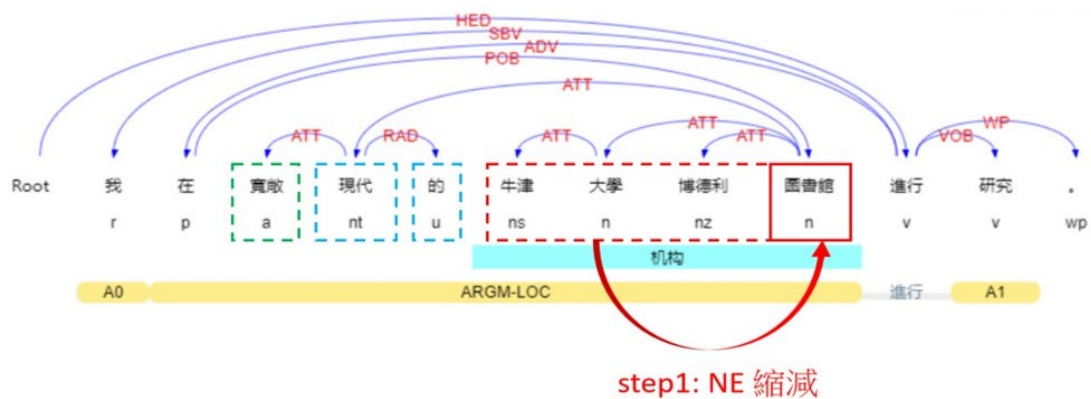
而在資料集標註部分，此次資料集目的在於收集和標記出每個句子中(不處理跨句的組合)所有修飾關係的詞彙組合，以供神經網絡模型做訓練或有利知識庫的收集，而此兩份語料因為語料特性在收集方式略有不同。在哈工大語料部分採用原始哈工大語法分析器使用之語料，而此份語料包含句子中所有具有依存關係(dependency)的詞彙組合，因此在標記上的方式則是根據此依存關係的資訊找出所有可能的修飾關係。而在小學數學語料部分，由於原始語料並沒有提供標準之依存關係的資訊，因此在收集上採用哈工大所開發之語法分析器來找出所有的依存關係，因此在流程上可能有些許的不同。而此次收集修飾關係部分，根據依存關係的資訊本實驗採用自動標註的方式，即根據哈工大依存分析樹所提供的依存關係，以人工檢視的方式，找出依存關係和本實驗關係類別的對應，並產生對應表，以此來做快速且大量的標註，對應表如附錄表格 A.1 所示。而此處以圖片 3.7 來說明自動標註的情況，在此例中其根據依存分析樹的資訊(圖片上方)，透過對應表進行轉換，將依存關係轉換成修飾關係類別並標註其句子中所有的修飾關係(圖片下方)，藉此來收集句子中所有的修飾關係組合。另外，句子中命名實體

(Named Entity, NE)的內部的組成，非本次修飾關係組合的收集對象，因此此處使用哈工大的依存分析器(parser)中的命名實體識別功能找尋命名實體，並將句子進行 NE 的縮減，縮減方式為將命名實體內部的詞彙以內部最後一個詞(通常可以代表整個命名實體的意思)統一做取代，將整個命名實體視為一個單一的詞彙以尋找外部詞彙和命名實體的修飾關係，如圖片 3.8 說明。



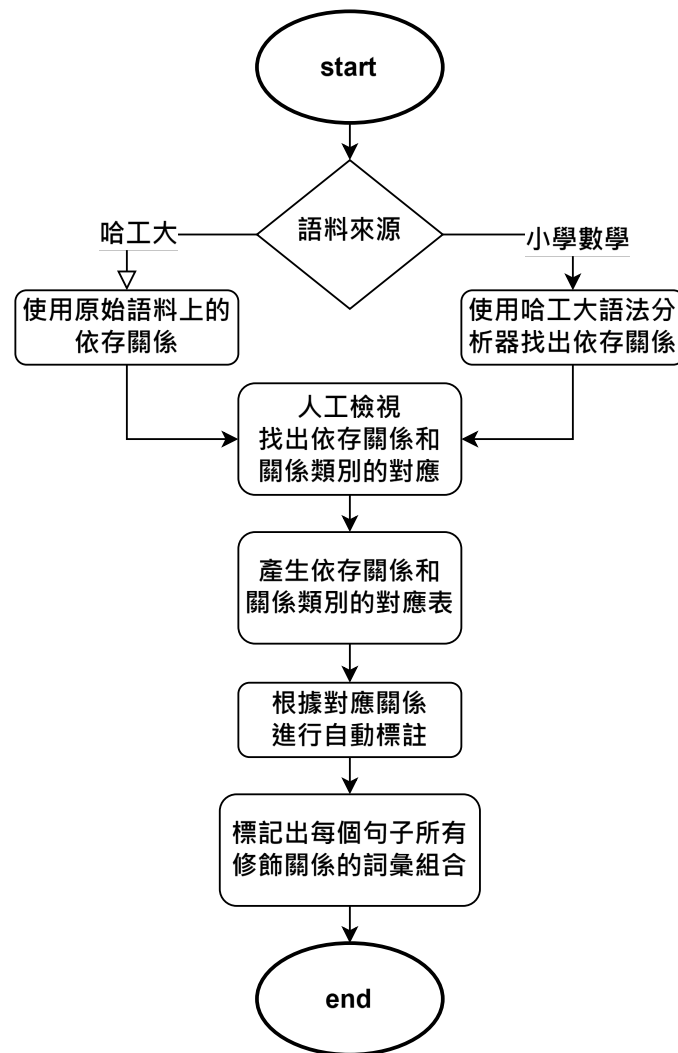
圖片 3.7：依存分析樹(上)和修飾關係連線圖(下)





圖片 3.8：命名實體處理

而在本實驗之關係類別中，亦需預測不具有修飾關係之組合，其類別為 False，而此類別沒辦法用語料所提供的依存關係取得，因此此類別之收集方法較為特殊，將在下一個小節另外做說明。而此收集修飾關係組合的整體流程整理如圖片 3.9。



圖片 3.9：收集修飾關係組合整體流程

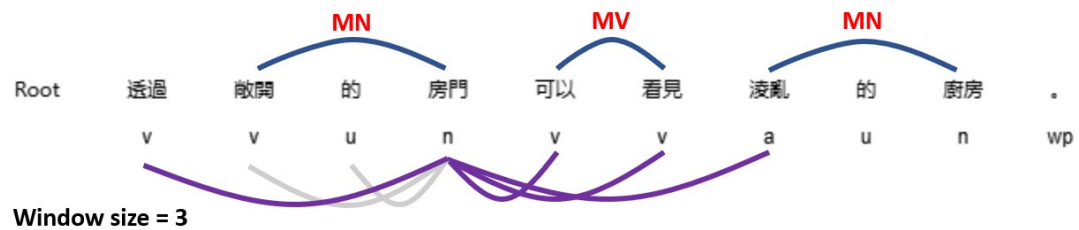
### 3.7.2 收集非修飾關係組合方法

而在收集非修飾關係組合，也就是關係標籤為 False 的組合，其方法為根據哈工大對中文句子的依存關係樹所剖析的詞性，先找出剖析詞性為名詞(n)和動詞(v)的部分。並以這些詞彙當作中心點取前後範圍內的其他詞彙和其組合成非修飾關係組合的詞彙組合，以此來收集非修飾的關係組合。而選取名詞、動詞為中心點的原因在於目前所收納的修飾關係中不外乎都是修飾名詞或動詞的關係，例如：MN、NM、PN 修飾名詞；MV、VM 修飾動詞。因此在詢問句子中的所有

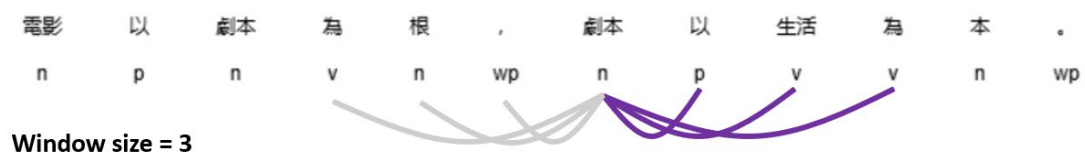
的修飾關係時，也都是以名詞和動詞出發去詢問和它的修飾關係，故此處以名詞和動詞作為中心點去收集範圍內的非修飾關係也較符合詢問時的情境。而中心點範圍附近的詞彙組合收集，則會設定一個固定的範圍(window size)，以此來排除過遠距離的組合。而在選取範圍部分(window size)，此 window size 功用是在於選取中心詞(名詞或動詞)前後範圍內可能和中心詞的詞彙組合，原因在修飾關係中有前面修飾以及後面修飾關係，例如：MN、MV、PN 為前面修飾關係；NM、VM 為後面修飾關係，故此 window size 的選取為中心詞前後固定距離內的詞彙組合。

而在選取中心詞前後 window size 內的詞彙組合並非所有都是非修飾關係(False)，其中還可能包括修飾關係和需要過濾掉的組合，另外命名實體內部的詞彙組合也不予以收集，僅根據命名實體縮減完的句子選取可能的非修飾關係組合。而以下將分別對需要過濾掉的組合進行說明。首先，因為選取中心詞前後 window size 內的詞彙組合目的在於收集非修飾關係的組合，因此必須先將修飾關係組合予以去除。以圖片 3.10 為例，假設以「房門」為中心詞前後提取 window size 範圍內的非修飾關係組合，而其中(敞開, 房門)為修飾關係所以不收入為非修飾關係的組合。再來，是去除跨句的組合，以「，」、「。」、「?」、「!」、「；」等等標點符號相隔兩個詞彙即視為跨句的情況，而此類型的組合不予以收集。如圖片 3.11 為去除跨句的組合的範例，在此例以中心詞為「劇本」為例，其詞性為 n，透過選取中心詞前後 window size 內的詞彙，在此處 window size 訂為 3 即是取中心詞前面 3 個斷詞和後面 3 個斷詞的組合，但因為前面 window size 範圍的詞彙有跨句的情況(有「，」相隔)，因此(為, 劇本)、(根, 劇本)、(，, 劇本)這三組合予以去除。接者，虛詞和中心詞的組合也需予以去除，因為在實際詢問場景中虛詞部分會先去除不會做詢問，而此處虛詞判斷標準為哈工大詞性標註為 u 者(詞性 u 在哈工大語法分析器中定義為助詞)。實際例子如圖片 3.12 所示，在此例中「談判」為中心詞，而「的」因為詞性為 u 屬於虛詞，因此(的, 談判)這個

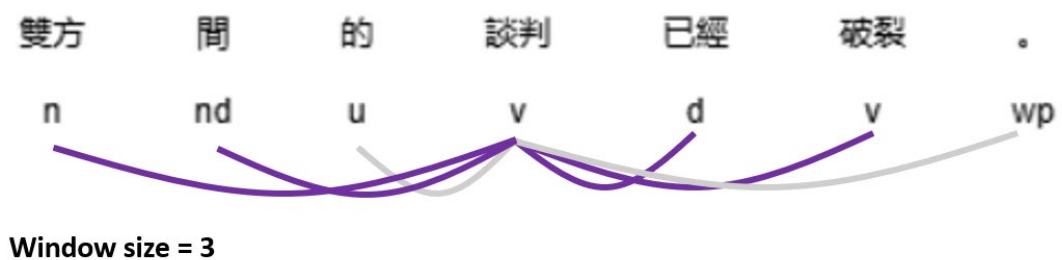
詞彙組合也不予以收集。而根據上述的收集方式，最後則是隨機抽取和修飾關係相同數量之非修飾關係當作最後的 False 收集結果，整體收集非修飾關係組合流程統整如圖片 3.13 所示。



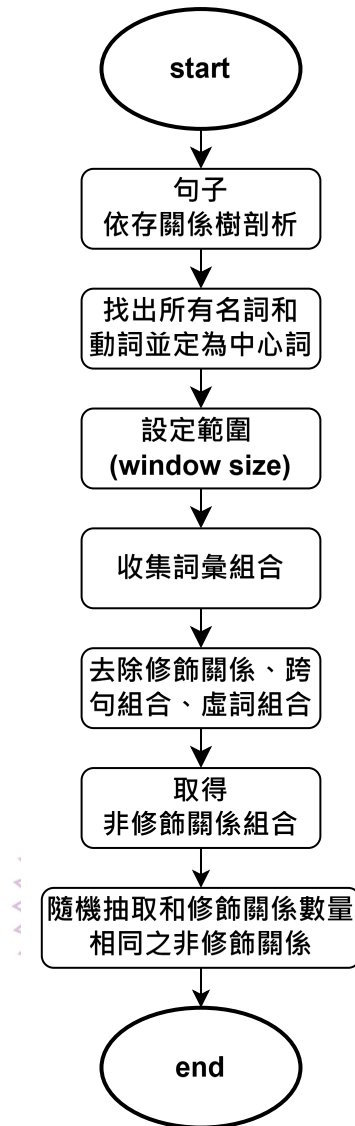
圖片 3.10：去除修飾關係組合



圖片 3.11：去除跨句的組合



圖片 3.12：去除虛詞的組合



圖片 3.13：收集非修飾關係組合整體流程

### 3.7.3 關係組合之資料分布

在此小節描述關係組合之資料分布，資料標籤分布如表格 3.6、表格 3.8，表格中包含了各個 label 的數量以及其在修飾關係 (True) 中的百分比，而表格 3.7、表格 3.9 在可以看到修飾關係 (True) 和非修飾關係(False)占全部數據的百分比，以下將分別針對哈工大語料和小學數學語料之資料標籤分布做說明。首先，可以看到在哈工大語料和小學數學語料中修飾關係數量(True)和非修飾關係

(False)各占 50%。另外表格 3.6 為哈工大語料之資料標籤分布，其中 MN 標籤最多占了約 60%，其次是 MV 占了約 33%，剩下其他類別加總約占 5~6%。而表格 3.8 為小學數學語料之資料標籤分布，其占比最高為 MV 類別約為 46%，其次是 MN 約為 27%，第三是 VM 約占 25%，其餘標籤占較小之比例。而根據表格 3.6、表格 3.8，可以發現兩個語料的資料標籤分布略有不同，此可能跟語料特性不同有關。另外在切分資料集部分，各個子資料集中資料標籤分布均和表表格 3.6、表格 3.8 維持一致。

Label	Count	Percentage
MN	85405	60.44%
MV	47638	33.71%
NM	65	0.05%
PN	149	0.11%
VM	8042	5.69%
Total	141299	100.00%

表格 3.6：哈工大資料集的 label 分布

Label	Count	Percentage
TRUE	141299	50.00%
FALSE	141299	50.00%
Total	282598	100.00%

表格 3.7 哈工大資料集的修飾、非修飾關係分布



Label	Count	Percentage
MN	14106	27.38%
MV	24123	46.83%
NM	22	0.04%
PN	134	0.26%
VM	13126	25.48%
Total	51511	100.00%

表格 3.8：小學數學資料集的 label 分布

Label	Count	Percentage
TRUE	51511	50.00%
FALSE	51511	50.00%
Total	103022	100.00%

表格 3.9 小學數學資料集的修飾、非修飾關係分布

### 3.8 神經網路模型之訓練方法

在此節中將介紹本實驗之神經網路模型的訓練方法，此處將分別針對參數的更新、訓練的方式、損失函數(Loss function)、優化器(Optimizer)等等的選擇做介紹。

首先參數更新部分此次研究選擇凍結(Freezing)預訓練模型的方式，凍結預訓練模型的參數更新，僅更新分類模型的參數部分。此好處在於能夠減少計算資源消耗、加速訓練過程，因為僅更新分類模型的參數，可以大幅減少計算量，節省訓練時間和資源。另外預訓練模型在大量數據上訓練，已經學習到豐富的特徵和知識，透過凍結這些參數可以保留這些知識，並將其應用到新任務中。而在訓練部分則是使用提前停止(early stopping)的方式，可以避免不必要的訓練，節省

計算資源和時間，並取出效果最佳的模型。另外也可以防止模型在訓練數據上過度擬合，以此來提高模型在數據上的泛化能力。而在損失函數部分使用交叉熵(cross entropy)，來計算預測的機率分布和理想機率分布的距離。而優化器(optimizer)部分使用 Adam(Adaptive Moment Estimation)，此優化器結合了動量(Momentum)和 RMSProp 的優點。動量的概念可以避免訓練過程中卡在 Critical Point 的情況，例如局部最小值(local min)或鞍點(saddle point)。而 RMSProp 的優點除了能夠根據梯度(Gradient)情況來調整 learning rate 外，還可以避免梯度變化差異過大所造成的問題。透過 Adam 的使用，其具有自適應學習率(Adaptive Learning Rates)的特性，能夠自動設定各參數之學習率，並自動調整學習率之大小。

### 3.9 神經網路模型參數設定

本研究神經網路所使用之預訓練模型架構及使用版本和超參數(hyper-parameters)設定如表格 3.10 所示。

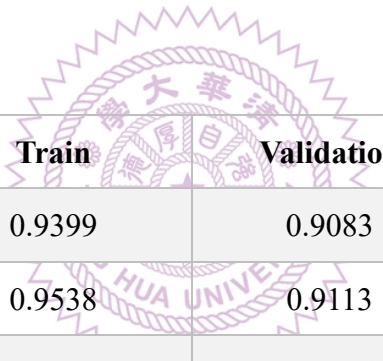
<b>Transformer 架構</b>	12-layer, 768-hidden, 12-heads, 110M parameters
<b>預訓練模型版本 (Pretrained model)</b>	RoBERTa-wwm-ext <sub>base</sub>
<b>損失函數(Loss function)</b>	交叉熵(Cross-Entropy)
<b>優化器(Optimizer)</b>	Adam
<b>學習率 (learning rate)</b>	$1 \times 10^{-4}$
<b>批次大小 (batch size)</b>	32
<b>Training epochs</b>	200

表格 3.10：實驗參數設定

## 第四章 實驗與結果討論

### 4.1 模型效果

表格 4.1、表格 4.2 分別為哈工大語料和小學數學語料效果最佳的模型之數據，包含模型之準確率(Accuracy)、精確率(Precision)、召回率(Recall)以及 F1-score。而表格 4.3、表格 4.3 分別為哈工大語料、小學數學語料模型效果比較表。此表中所使用到知識庫(KB)部分為經過篩選，選取詞彙組合為類型一：單一修飾關係類別後的結果，透過知識庫篩選類別一的詞彙組合，綜合預測效果可以達到最佳，說明在 4.2 小節：知識庫篩選比較。而透過此表可以發現，綜合知識庫和神經網路的預測效果是可以勝過單獨使用神經網路或是知識庫的方式，其結合的預測效果能夠達到最佳。



	Train	Validation	Test
Accuracy	0.9399	0.9083	0.9073
Precision	0.9538	0.9113	0.8400
Recall	0.9591	0.9156	0.8774
F1-score	0.9554	0.9132	0.8513

表格 4.1：哈工大語料整體模型效果

	Train	Validation	Test
Accuracy	0.9838	0.9681	0.9669
Precision	0.9869	0.9771	0.9614
Recall	0.9931	0.9829	0.9664
F1-score	0.9899	0.9799	0.9639

表格 4.2：小學數學語料整體模型效果

	<b>Train</b>	<b>Validation</b>	<b>Test</b>
<b>KB+NN Accuracy</b>	<b>0.9399</b>	<b>0.9083</b>	<b>0.9073</b>
<b>NN Accuracy</b>	0.9120	0.9058	0.9052
<b>KB Accuracy</b>	0.9312	0.6826	0.6809

表格 4.3：哈工大語料模型效果比較表

	<b>Train</b>	<b>Validation</b>	<b>Test</b>
<b>KB+NN Accuracy</b>	<b>0.9838</b>	<b>0.9681</b>	<b>0.9669</b>
<b>NN Accuracy</b>	0.9767	0.9666	0.9652
<b>KB Accuracy</b>	0.9050	0.8219	0.8193

表格 4.4：小學數學語料模型效果比較表

## 4.2 知識庫篩選比較

如 3.4 小節：關係預測整體流程所述，目前關係類別預測的方法架構為知識庫 (Knowledge base, KB) 和神經網路 (Neural Network, NN) 進行綜合的預測 (KB+NN)。其運作方式為輸入一個中文句子以及當中的一個詞彙組合，若知識庫中存在此詞彙組合，則會輸出知識庫預測的結果；若知識庫不存在此詞彙組合，則交由神經網路進行最後的預測。另外在表格 4.5、表格 4.6 中有單獨使用知識庫的方法，此處先解釋此方法的預測方式，其預測方式為判斷此詞彙組合是否存在於知識庫中，若存在則預測其記錄的關係類別；若不存在則預測為非修飾關係 (False)。

而在此兩個方法的綜合預測當中，知識庫還需進行適當的篩選，以便篩選出品質較高的詞彙組合，達到最好的綜合效果。如表格 4.5、表格 4.6 所示，其分

別為哈工大和小學數學語料在知識庫未篩選情況下的模型效果比較表，若知識庫的詞彙組合未經過適當的篩選，其綜合效果可能會比單獨使用神經網路模型或是知識庫模型還差。如表格 4.5、表格 4.6 中均可以看出在訓練資料集(train set)效果最好者為單純用知識庫的模型，而在驗證資料集(validation set)以及測試資料集(test set)效果最好者都為單純使用神經網路的模型，其綜合效果比單獨使用神經網路模型或是知識庫模型還差。因此如何篩選知識庫的詞彙組合以達到最好的綜合預測效果為此小節研究的重點。

	Train	Validation	Test
<b>KB+NN Accuracy</b>	0.9339	0.9051	0.9035
<b>NN Accuracy</b>	0.9120	<b>0.9058</b>	<b>0.9052</b>
<b>KB Accuracy</b>	<b>0.9775</b>	0.7304	0.7290

表格 4.5：哈工大語料模型效果比較表(知識庫未篩選)

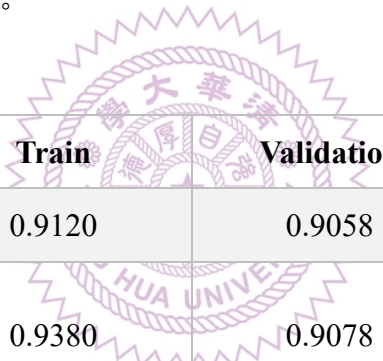
	Train	Validation	Test
<b>KB+NN Accuracy</b>	0.9747	0.9622	0.9601
<b>NN Accuracy</b>	0.9767	<b>0.9666</b>	<b>0.9652</b>
<b>KB Accuracy</b>	<b>0.9834</b>	0.9021	0.8981

表格 4.6：小學數學語料模型效果比較表(知識庫未篩選)

而根據 3.6 小節：知識庫的建立，可以得知詞彙組合分類主要有四種，分別是類型一：單一修飾關係類別、類型二：修飾/非修飾關係類別、類型三：多修飾關係類別、類型四：純非修飾關係類別。而表格 4.7、表格 4.8 為知識庫(KB)篩

選各種類型結合神經網路(NN)進行綜合預測的結果比較，而其中類型四因其預測類別為非修飾關係(False)在知識庫中不予以紀錄，因此不予以進行比較。

而在效果比較部分，如表所示在以下兩個語料中，知識庫中篩選出類型一的詞彙組合結合神經網路的綜合效果最佳，高於單純使用神經網路的模型(NN (No KB))以及其他的篩選類型結合神經網路的模型。其原因在於類型一、二、三之間的差異，類型一為詞彙組合僅出現一種關係類別的情況；而類別二、三為詞彙組合出現多於一種關係類別的情況。所以透過此類別之間的差異，可以得出以下的結論，在知識庫結合神經網路的方法中，若知識庫只收集類型一：單一修飾關係類別，也就是收集只出現一種關係類別的情況，等同是讓知識庫只記錄其確定的詞彙組合，而剩下可能出現多於一種關係類別的詞彙組合則交由神經網路模型做預測其綜合效果是最佳的。



	<b>Train</b>	<b>Validation</b>	<b>Test</b>
<b>NN (No KB)</b>	0.9120	0.9058	0.9052
<b>KB (類型 1,2,3) + NN</b>	0.9380	0.9078	0.9057
<b>KB(類型 1,3) + NN</b>	0.9393	0.9078	0.9059
<b>KB(類型 1) + NN</b>	<b>0.9399</b>	<b>0.9083</b>	<b>0.9073</b>

表格 4.7：哈工大語料模型知識庫類型篩選效果比較表



	<b>Train</b>	<b>Validation</b>	<b>Test</b>
<b>NN (No KB)</b>	0.9767	0.9666	0.9652
<b>KB (類型 1,2,3) + NN</b>	0.9804	0.9673	0.9661
<b>KB(類型 1,3) + NN</b>	0.9837	0.9680	0.9668
<b>KB(類型 1) + NN</b>	<b>0.9838</b>	<b>0.9681</b>	<b>0.9669</b>

表格 4.8：小學數學語料模型知識庫類型篩選效果比較表

透過上述的比較，可以得知知識庫中篩選出類型一的詞彙組合結合神經網路的方式效果最佳。而表格 4.9、表格 4.10 為此效果最佳模型和單獨使用神經網路或是單獨使用知識庫的比較。透過此表可以發現，知識庫經過適當的篩選後，其綜合知識庫和神經網路的預測效果是可以勝過單獨使用其中任一種的方式，有別於先前表格 4.5、表格 4.6 的結果，其結合的預測效果能夠達到最佳。

	<b>Train</b>	<b>Validation</b>	<b>Test</b>
<b>KB+NN Accuracy</b>	<b>0.9399</b>	<b>0.9083</b>	<b>0.9073</b>
<b>NN Accuracy</b>	0.9120	0.9058	0.9052
<b>KB Accuracy</b>	0.9312	0.6826	0.6809

表格 4.9：哈工大語料模型效果比較表(知識庫篩選類型一)

	<b>Train</b>	<b>Validation</b>	<b>Test</b>
<b>KB+NN Accuracy</b>	<b>0.9838</b>	<b>0.9681</b>	<b>0.9669</b>
<b>NN Accuracy</b>	0.9767	0.9666	0.9652
<b>KB Accuracy</b>	0.9050	0.8219	0.8193

表格 4.10：小學數學語料模型效果比較表(知識庫篩選類型一)

### 4.3 人工驗證自動標註之正確率

此處人工驗證自動標註的目的在於，此研究之資料集產生是採用自動標註產生大量且快速的標註。而自動標註方法如 3.7.1 小節：收集修飾關係組合方法所述，是採用哈工大的依存關係樹經過對應表轉換成此研究的修飾關係類別而來，因此此處需要經過人工驗證以驗證自動標註的準確度。

而此人工驗證方法為隨機抽取句子進行人工檢視，若正確率很高則代表此自動標註方法有效。因此目前做法是分別從哈工大、小學數學語料中各隨機抽取 200 個詞彙組合，人工檢視其自動標註的結果是否正確，以此來做驗證。而目前僅抽取是修飾關係(True)的部分，因為此處主要在於驗證自動標註的準確度，而自動標註主要負責修飾關係的收集，因此只收集修飾關係，而非修飾關係(False)則不在抽取範圍中。而目前人工驗證自動標註的正確率如表格 4.11、表格 4.12，在哈工大語料中隨機抽取的 200 個詞彙組合當中，正確標註的有 198 個，正確率為 99%；而在小學數學語料中隨機抽取的 200 個詞彙組合當中，均為正確標註，正確率為 100%。

	哈工大語料
正確率	99.00% (198/200)

表格 4.11：哈工大語料人工驗證自動標註正確率

	小學數學語料
正確率	100% (200/200)

表格 4.12：小學數學語料人工驗證自動標註正確率

## 4.4 錯誤分析

表格 4.13、表格 4.14 分別為哈工大、小學數學語料錯誤分布統計表，目前主要分析知識庫(篩選類型一)結合神經網路模型進行綜合預測之結果。而目前分析方法為從 30 個預測錯誤的詞彙組合中進行錯誤分析，而其錯誤類型分布如下，錯誤類型主要有兩種預測錯誤及標註錯誤。在哈工大語料錯誤分布如表格 4.11，預測錯誤為 26 個，約佔 87%；而標註錯誤為 4 個，約佔 13%，數量以預測錯誤較多；而在小學數學語料錯誤分布如表格 4.12，全部錯誤均為預測錯誤，沒有標註錯誤的部分。

錯誤類型	預測錯誤	標註錯誤	錯誤數量
數量	26	4	30
比例	86.66%	13.33%	100%

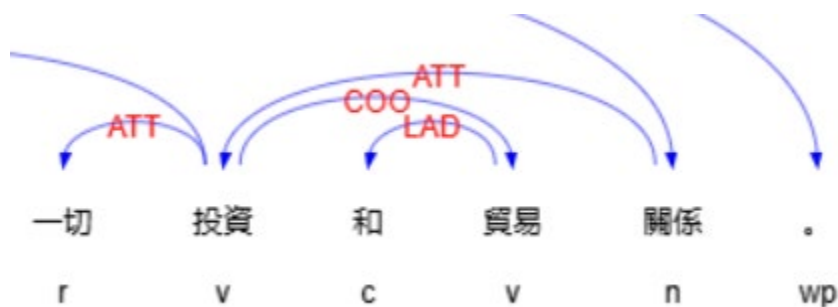
表格 4.13：哈工大語料錯誤分布統計表

錯誤類型	預測錯誤	標註錯誤	錯誤數量
數量	30	0	30
比例	100%	0%	100%

表格 4.14：小學數學語料錯誤分布統計表

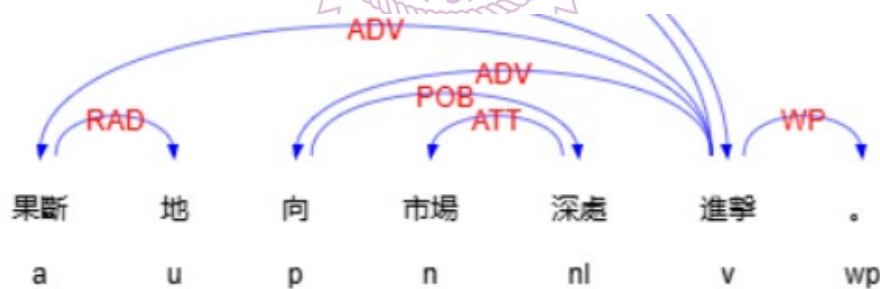
而根據觀察實際錯誤案例，在標註錯誤部分，主要由兩個錯誤來源所構成，其分別為並列結構問題及 pattern 問題所導致(pattern 為哈工大依存關係的表示形式，附錄 A：哈工大依存關係對應修飾關係分類表有詳細說明)，以下將分別進行說明。首先是並列結構問題部分，如圖片 4.1 的例子，假設欲判斷在句中(貿易，關係)是否是修飾關係，依人為判斷應為 MN 修飾關係，然而因哈工大的依存結構因素，其會將(投資，貿易)先視為並列結構，再統一和「關係」判斷其依

存關係，因此(貿易，關係)並沒有直接的依存關係連線，故導致標註錯誤。而此問題的解決方式，可能需要透過增加新的修飾關係或特別處理此結構得以解決。



圖片 4.1：並列結構問題案例之哈工大依存分析樹

而 pattern 問題，則是依存關係和修飾關係的對應收集的不完全所導致的自動標註缺失。而其案例說明如圖片 4.2，在此句中，(市場，深處)此詞彙組合應為 MN 關係，其依存關係轉換成 pattern 形式為 n-nl-ATT，然而此 pattern 目前尚未加入到自動標註中，因此導致標註錯誤。而此類型問題有待後續增加更多人力來進行人工判斷，以增加更多自動標註的 pattern 得以解決。



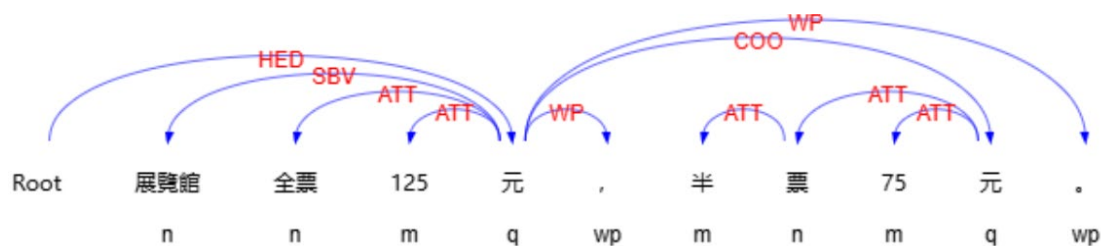
圖片 4.2：pattern 問題案例之哈工大依存分析樹

## 4.5 整體模型方法和哈工大對應表方法的比較

此小節為本研究之整體預測方法(即知識庫結合神經網路模型)和直接使用哈工大語法分析器結合對應表方法(簡稱哈工大對應表方法)進行預測的比較。因

透過直接使用哈工大對應表方法，先對句子進行哈工大依存關係的分析再透過對應表轉換成收集的修飾關係，也可以預測詞彙組合之間關係，故比較以上兩種方法在實際應用場景上的差異有其必要性。因此在此小節主要探討本研究之整體預測方法之目的並說明其優勢，並且解釋為何此方法勝過直接使用哈工大對應表方法進行預測的方式。

此處以實際例子來進行說明。假設有一中文句子如圖片 4.3 所示，欲詢問(展覽館, 全票)此詞彙組合之關係。此詞彙組合在人為判斷下此處應為 MN 之修飾關係，但由於此處哈工大語法分析器可能誤判的情況發生，導致判斷不出此詞彙組合具有依存關係，因此可能會導致產生預測錯誤的結果。而本研究整體預測方法解決此問題的方式主要有以下兩點： 1. 假設此詞彙組合存在於知識庫當中並且記錄為修飾關係，則在此處即可正確的判斷，而不受哈工大語法分析錯誤的影響。 2. 若此處標註是以人工的方式進行標註並交由模型訓練，則此處若模型訓練得當，可正確預測此組合的修飾關係，而不受哈工大語法分析錯誤的影響。此兩種方式均可以解決直接使用哈工大對應表方法在圖片 4.3 例子中所遇到的問題(即哈工大語法分析器的錯誤)，故使用本研究之整體預測方法有其必要性，其效果並非直接使用哈工大對應表方法進行預測所能達到。



圖片 4.3：哈工大語法分析器找不到依存關係之實際案例



# 第五章 結論與未來展望

## 5.1 結論

自然語言的複雜性增加了閱讀者準確把握文本內容的難度，也影響了系統處理文本的效率。然而我們可以將複雜的語言結構簡化為核心語義概念。這種方法不僅能夠保留文本的關鍵信息，還能大幅減少需要處理的數據量。而本研究之結果可以應用於中文句子之簡化法，其方法為找出句子中的修飾關係將複雜句縮簡為簡單句，藉此還原主語(Subject, S)、謂語(Verb, V)、賓語(Object, O)的結構。而本研究提出了一種基於知識庫與神經網路模型結合的中文詞彙組合關係分類方法，並成功應用於句子簡化及修飾關係提取中。此方法透過找出兩兩詞彙組合的修飾關係，能夠找出句子中所有的修飾關係搭配詞，提供給後續語法分析器(parser)做使用，輔助其產生句子的依存分析樹。

此外，本研究也提出了詞彙修飾關係之資料集。在資料集的製作方面，經過人工驗證，自動標註方法的準確率也都達到了99%以上，證明了此自動標註方法的有效性和可靠性。而實驗結果部分，在知識庫的建立與篩選方面，透過對詞彙組合進行類別篩選，特別是僅收集單一修飾關係類別的詞彙組合，能夠顯著提升模型的綜合預測效果。這證明了高品質詞彙組合對應關係的篩選對於提升預測準確度的重要性。

## 5.2 未來展望

在本研究的基礎上，未來有多個方向可以進一步提升中文句子簡化法的方法與應用。首先，現有方法是透過兩兩尋找詞彙間的修飾關係，未來可以改進為一次性找出句子中所有的修飾關係，這樣不僅可以提高處理效率，還能更全面地捕捉句子的修飾結構。其次，可以考慮增加修飾關係的類別，目前研究中主要收集了五種修飾類別，未來可以擴展這些類別，以涵蓋更多樣化的修飾關係，從而提



升模型的預測能力和應用範圍。

而此外語料庫的擴展也是未來的重要方向之一。現有語料庫主要來自哈工大語料和小學數學語料，未來可以引入更多不同領域和風格的文本資料，如新聞、論壇網站等等來源，以提升模型的泛化能力和適應性。另外將此方法應用於其他語言也是一個值得探索的方向，可以驗證其跨語言的適用性和效果，並進一步改進模型以適應不同語言的特性。最後，這種方法未來不僅限於靜態文本的處理，後續還可以應用於語言生成。可以透過修飾關係將簡單句增加為複雜句，利用收集的修飾關係對句子進行生成、擴充。

總結來說，未來的研究可以通過一次性找出句子中所有修飾關係、增加修飾關係類別、擴展語料庫範圍、應用於其他語言以及語言生成等多個方向，進一步提升中文句子簡化法之實際效果和應用價值。



## 參考文獻

- [1] "處理自然語言的簡化法," Institute of Information Science, Academia Sinica, <https://ipitt.sinica.edu.tw/shares/929>, May 2023 (Jul. 17, 2024).
- [2] "自然語言簡化法 (Reduction)," Taiwan Bioinformatics Institute, <http://www.tbi.org.tw/enews/TMBD/Vol38.html>, October 2021 (Jul. 17, 2024).
- [3] F. Alva-Manchego, C. Scarton, and L. Specia, "Data-Driven Sentence Simplification: Survey and Benchmark," *Computational Linguistics*, vol. 46, no. 1, pp. 135–187, 2020.
- [4] M. T. Nguyen, C. M. Bui, D. T. Le, and T. L. Le, "Sentence compression as deletion with contextual embeddings," in *International Conference on Computational Collective Intelligence*, Cham: Springer International Publishing, 2020, pp. 427–440.
- [5] K. Filippova, "Multi-sentence compression: Finding shortest paths in word graphs," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 2010, pp. 322–330.
- [6] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 615–620.
- [7] A. Madotto, C. Wu, and P. Fung, "Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems," in *Proceedings of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1468–1478.
- [8] C. Xiong, R. Power, and J. Callan, "Explicit semantic ranking for academic search via knowledge graph embedding," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1271–1279..

- [9] X. Han, et al., "More data, more relations, more context and more openness: A review and outlook for relation extraction," *arXiv preprint arXiv:2004.03186*, 2020.
- [10] H. Wang, K. Qin, R. Y. Zakari, et al., "Deep neural network-based relation extraction: an overview," *Neural Comput & Applic*, vol. 34, pp. 4781–4801, 2022.
- [11] P. Zhou, S. Zheng, J. Xu, Z. Qi, H. Bao, and B. Xu, "Joint Extraction of Multiple Relations and Entities by Using a Hybrid Neural Network," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. NLP-NABD CCL 2017, Lecture Notes in Computer Science*, vol. 10565, M. Sun, X. Wang, B. Chang, and D. Xiong, Eds. Cham: Springer, 2017.
- [12] H. Peng, et al., "Learning from context or names? an empirical study on neural relation extraction," *arXiv preprint arXiv:2010.01923*, 2020.
- [13] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-Training With Whole Word Masking for Chinese BERT," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2021.
- [14] Y. Huang, et al., "D-BERT: Incorporating dependency-based attention into BERT for relation extraction," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 4, pp. 417–425, 2021..
- [15] J. Devlin, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [16] S. Wang, "The Survey of Joint Entity and Relation Extraction," in *Computing and Data Science. CONF-CDS 2021. Communications in Computer and Information Science*, vol. 1513, W. Cao, A. Ozcan, H. Xie, and B. Guan, Eds. Singapore: Springer, 2021.
- [17] M. Shardlow, "A survey of automated text simplification," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 1, pp. 58–70, 2014.

- [18] S. Štajner, K. C. Sheang, and H. Saggion, "Sentence simplification capabilities of transfer-based models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 12172-12180.
- [19] J. Clarke and M. Lapata, "Global inference for sentence compression: An integer linear programming approach," *J. Artif. Intell. Res.*, vol. 31, pp. 399-429, 2008.
- [20] K. Filippova and Y. Altun, "Overcoming the lack of parallel data in sentence compression," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1481-1491.
- [21] M. E. Califf and R. J. Mooney, "Relational learning of pattern-match rules for information extraction," in *Proceedings of CoNLL*, 1997.
- [22] K. Fundel, R. Küffner, and R. Zimmer, "RelEx—Relation extraction using dependency parse trees," *Bioinformatics*, vol. 23, no. 3, pp. 365-371, 2007.
- [23] M. Cui, et al., "A survey on relation extraction," in *Knowledge Graph and Semantic Computing. Language, Knowledge, and Intelligence: Second China Conference, CCKS 2017*, Chengdu, China, August 26–29, 2017, Revised Selected Papers 2, Springer Singapore, 2017, pp. 50-58.
- [24] G. Bekoulis, J. Deleu, T. Demeester, et al., "Joint entity recognition and relation extraction as a multi-head selection problem," *Expert Syst. Appl.*, vol. 114, pp. 34–45, 2018.
- [25] J. Lee, S. Seo, and Y. S. Choi, "Semantic relation classification via bidirectional LSTM networks with entity-aware attention using latent entity typing," *Symmetry*, vol. 11, no. 6, p. 785, 2019.
- [26] S. Wu and Y. He, "Enriching pre-trained language model with entity information for relation classification," *arXiv preprint arXiv:1905.08284*, 2019.
- [27] M. Eberts and A. Ulges, "Span-based joint entity and relation extraction with transformer pretraining," *arXiv preprint arXiv:1909.07755*, 2019.
- [28] C. Dong, et al., "A survey of natural language generation," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1-38, 2022.

- [29] Y. Safovich and A. Azaria, "Fiction sentence expansion and enhancement via focused objective and novelty curve sampling," in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 2020, pp. 835-843.
- [30] T. Iqbal and S. Qureshi, "The survey: Text generation models in deep learning," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 2515-2528, 2022.
- [31] Y. Zhang, Y. Wang, and J. Yang, "Lattice LSTM for Chinese sentence representation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1506-1519, 2020.
- [32] J. Devlin, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.



# 附錄

## A. 哈工大依存關係對應修飾關係分類表

附錄表格 A.1 為哈工大依存關係對應修飾關係分類表，而表中第一欄之依存關係以 Pattern 來表示，而對應的修飾關係則以 Label 來做表示，此表的目的是在於提供自動標註時的對應關係。而自動標註的流程如 3.7.1 小節：收集修飾關係組合方法所述，其流程為透過哈工大的語法分析器分析句子的依存關係後，再根據此對應表進行自動標註，最後產生整句的修飾關係。

此表僅收集對應關係為修飾關係的部分，因自動標註僅標註修飾關係的組合，故所有的依存關係(Pattern)對應修飾關係(Label)如下表所示。另外，表中 pattern 部分為哈工大的語法分析器的分析結果轉換為 dependent-head-relation 的表示方式。因為此 pattern 不考慮方向性，而 label 中具有方向性(如 MN 和 NM 的區別)，故表中 pattern 對應的 MN、MV 標籤會根據 dep、head 的詞彙所在的相對位置判斷是否轉換成 NM、VM。例如有一個句子「全隊進攻快速。」，其中有一詞彙組合(進攻, 快速)，在經過哈工大的語法分析器分析後，其分析結果「進攻」為 head、「快速」為 dependent、兩者的關係為 CMP，因此其 pattern 結果為 d-v-CMP。在附錄表格 A.1 中 d-v-CMP 其對應的 Label 為 MV，然而在此句子中「進攻」在前面、「快速」在後方，因此將此組合標示成 VM。

Pattern	Label
n-n-ATT	MN
d-v-ADV	MV
v-n-ATT	MN
a-n-ATT	MN
v-v-ADV	MV



v-v-CMP	MV
b-n-ATT	MN
nt-v-ADV	MV
nd-v-ADV	MV
j-n-ATT	MN
n-n-SBV	MN
p-n-ATT	PN
n-v-ATT	MN
a-v-CMP	MV
n-n-FOB	MN
i-n-ATT	MN
a-n-ADV	MN
a-v-ADV	MV
i-v-ADV	MV
n-n-ADV	MN
v-v-ATT	MN
d-v-ATT	MN
v-v-RAD	MV
d-n-ATT	MN
n-v-CMP	MV
nl-v-ADV	MV
r-v-ATT	MN
m-v-CMP	MV
b-n-ADV	MN

b-v-ADV	MV
d-v-CMP	MV
nh-v-ATT	MN
n-b-COO	MN
i-n-ADV	MN
u-n-ATT	MN
b-n-SBV	MN
j-n-SBV	MN
k-n-ATT	MN
k-n-SBV	MN
n-n-RAD	MN
v-n-LAD	MN

附錄表格 A.1：哈工大依存關係對應修飾關係分類表

## B. 知識庫頻率篩選

根據 4.2 小節：知識庫篩選比較，其知識庫是透過辭彙組合的類別所進行的篩選。而此處還有另外一種篩選知識庫中辭彙組合的方式，是透過辭彙組合的頻率進行篩選。此處假設高頻率的辭彙組合其重要性應該高於低頻率的辭彙組合，因此此處實驗透過頻率篩選辭彙組合的方式，在知識庫中收集高頻的辭彙組合，觀測其是否會提升整體預測的效能。

以附錄表格 B.1、附錄表格 B.2 來說明，此兩個表格分別為哈工大和小學數學語料模型的知識庫頻率篩選效果比較表，圖中的數據均為知識庫和神經網路綜合預測的數據。在下表中，第一欄為頻率篩選的門檻，將低於設定數值的辭彙組合從知識庫中刪除。首先附錄表格 B.1：哈工大語料模型(NN+KB)知識庫頻率篩選效果比較表的結果可以觀察頻率篩選值設定在 0 時在訓練集中可以達到最好的效果，而設定篩選值為 10 能在驗證集和測試集達到最好的效果。而在附錄表

格 B.2：小學數學語料模型(NN+KB)知識庫頻率篩選效果比較表中，頻率篩選值設定在 500 甚至不使用知識庫時在訓練集中可以達到最好的效果，而在驗證集和測試集部分分別設定為 20 和 40 時效果最佳。

Min Frequency	Train	Validation	Test
<b>0</b>	<b>0.9339</b>	0.9051	0.9035
<b>10</b>	0.9129	<b>0.9065</b>	<b>0.9061</b>
<b>20</b>	0.9124	0.9060	0.9058
<b>50</b>	0.9122	0.9060	0.9054
<b>100</b>	0.9120	0.9058	0.9053
<b>no KB</b>	0.9120	0.9058	0.9052

附錄表格 B.1：哈工大語料模型(NN+KB)知識庫頻率篩選效果比較表

Min Frequency	Train	Validation	Test
<b>0</b>	0.9747	0.9622	0.9600
<b>20</b>	0.9750	<b>0.9675</b>	0.9640
<b>30</b>	0.9765	0.9674	0.9653
<b>40</b>	0.9765	0.9669	<b>0.9654</b>
<b>50</b>	0.9765	0.9666	0.9653
<b>100</b>	0.9765	0.9666	0.9652
<b>200</b>	0.9765	0.9665	0.9652
<b>500</b>	<b>0.9767</b>	0.9666	0.9652
<b>no KB</b>	<b>0.9767</b>	0.9666	0.9652

附錄表格 B.2：小學數學語料模型(NN+KB)知識庫頻率篩選效果比較表

而根據附錄表格 B.1、附錄表格 B.2 得出的結論，將頻率篩選中效果最好的模型與 4.2 小節類型篩選結果做比較，比較數據如附錄表格 B.3、附錄表格 B.4 所示。可以發現即使選取頻率篩選結果最好的模型和類別篩選做比較，類別篩選的效果還是勝過頻率篩選的結果。由此可以顯現相較於辭彙組合的頻率篩選，根據辭彙組合的類別進行篩選的結果，較能精準得篩選出高品質的辭彙組合，而由此小節也可以證明類別篩選的重要性，能夠輔助模型達到最好的預測效果。

	<b>Train</b>	<b>Validation</b>	<b>Test</b>
<b>NN (No KB)</b>	0.9120	0.9058	0.9052
<b>KB (類型 1,2,3) + NN</b>	0.9380	0.9078	0.9057
<b>KB (類型 1,3) + NN</b>	0.9393	0.9078	0.9059
<b>KB (類型 1) + NN</b>	<b>0.9399</b>	<b>0.9083</b>	<b>0.9073</b>
<b>KB (Min_freq = 0) + NN</b>	0.9339	0.9051	0.9035
<b>KB (Min_freq = 10) + NN</b>	0.9129	0.9065	0.9061

附錄表格 B.3：哈工大語料模型知識庫類型篩選和頻率篩選效果比較表

	<b>Train</b>	<b>Validation</b>	<b>Test</b>
<b>NN (No KB)</b>	0.9767	0.9666	0.9652
<b>KB (類型 1,2,3) + NN</b>	0.9804	0.9673	0.9661
<b>KB(類型 1,3) + NN</b>	0.9837	0.9680	0.9668
<b>KB(類型 1) + NN</b>	<b>0.9838</b>	<b>0.9681</b>	<b>0.9669</b>
<b>KB (Min_freq = 20) + NN</b>	0.9750	0.9675	0.9640
<b>KB (Min_freq = 40) + NN</b>	0.9675	0.9669	0.9654

附錄表格 B.4：小學數學語料模型知識庫類型篩選和頻率篩選效果比較表

## C. 知識庫和 Bigram 的比較

Bigram 是由兩個連續的詞組成的詞彙組合，屬於 n-gram 一種（n 表示詞的數量），是一種在自然語言處理（NLP）統計方法中常常使用的概念。而此處使用的 bigram，其收集方式將兩個語料經過哈工大的語法分析器的斷詞結果收集而成的。而其標註規則如附錄表格 C.1，此表為 bigram 詞性(Part of speech tag, POS-tag)組合和預測修飾關係(Label)的對應。此 POS-tag pair 為 bigram 中兩個詞彙的詞性，而表中 All 代表所有的詞性。

POS-tag pair	Label
All-n、n-n	MN
n-All	NM
All-v、v-v	MV
v-All	VM

附錄表格 C.1：bigram 詞性組合對應標註表

附錄表格 C.2、附錄表格 C.3 為知識庫和 bigram 的數量統計和集合比較表。以附錄表格 C.2 來看，若以聯集當作分母，交集比例約占 30%、KB 額外收集部分約占 10%、bigram 額外收集約占 60%。以附錄表格 C.3 來看，以聯集當作分母，交集比例約占 30%、KB 額外收集部分約占 20%、bigram 額外收集約占 50%。

	KB pair	Bigram pair	Union	Intersection	KB - Bigram (KB 額外收集)	Bigram - KB (Bigram 額外收集)
數量	80244	178513	198446	58806	21438	119195
比例			100%	29.63%	10.80%	60.06%

附錄表格 C.2：哈工大語料知識庫和 bigram 比較表



	<b>KB pair</b>	<b>Bigram pair</b>	<b>Union</b>	<b>Intersection</b>	<b>KB - Bigram (KB 額外收集)</b>	<b>Bigram - KB (Bigram 額外收集)</b>
<b>數量</b>	9375	15602	19029	5865	3510	9728
<b>比例</b>			100.00%	30.82%	18.45%	51.12%

附錄表格 C.3：小學數學語料知識庫和 bigram 比較表

而附錄表格 C.4、附錄表格 C.5 分別為哈工大、小學數學語料模型 bigram 頻率篩選效果比較表，數據的結果為知識庫和神經網路綜合的預測效果。而此處比較目的為觀察 bigram 的加入到既有的知識庫(未做類別及頻率篩選)中是否可以輔助知識庫做辭彙組合的修飾關係判斷，另外此處既有的知識庫未做類別及頻率篩選。而根據表中的結果可以得知，在不加入 bigram 到知識庫中的綜合預測效果是最好的。此顯現既有的知識庫收集方式，以足夠應付預測修飾關係的情況，bigram 的加入反而會影響到知識庫收集的品質，進而可能導致做出錯誤的判斷。

<b>Min Frequency</b>	<b>Train</b>	<b>Validation</b>	<b>Test</b>
<b>0</b>	0.8547	0.8378	0.8393
<b>30</b>	0.9324	0.9036	0.9016
<b>50</b>	0.9329	0.9043	0.9028
<b>100</b>	0.9336	0.9051	0.9034
<b>200</b>	<b>0.9339</b>	<b>0.9051</b>	<b>0.9036</b>
<b>No bigram</b>	<b>0.9339</b>	<b>0.9051</b>	<b>0.9036</b>

附錄表格 C.4：哈工大語料模型(NN+KB) bigram 頻率篩選效果比較表

Min Frequency	Train	Validation	Test
<b>0</b>	0.9545	0.9459	0.9400
<b>30</b>	0.9708	0.9594	0.9558
<b>50</b>	0.9740	0.9631	0.9592
<b>100</b>	0.9740	0.9631	0.9592
<b>200</b>	<b>0.9755</b>	<b>0.9647</b>	<b>0.9606</b>
<b>No bigram</b>	<b>0.9755</b>	<b>0.9647</b>	<b>0.9606</b>

附錄表格 C.5：小學數學語料模型(NN+KB) bigram 頻率篩選效果比較表

