HanTrans: An Empirical Study on Cross-Era Transferability of Chinese Pre-trained Language Model

(預訓練語言模型在漢語上的跨時代學習能力)

Chin-Tung Lin

Academia Sinica cindylin@iis.sinica.edu.tw

Wei-Yun Ma

Academia Sinica ma@iis.sinica.edu.tw

摘要

近年來預訓練的語言模型在自然語言處理 中蔚爲風潮,以BERT (Bidirectional Encoder Representations form Transformers) 模型爲代表,其掩碼語言建模 (maskedlanguage modeling, MLM) 被廣泛應用於 大型語言模型的預訓練,使得後續微調 (fine tuning) 後的模型即可在下游任務有 很好的表現。然而,在預訓練的語料中, 相比於簡體中文,繁體中文只佔了很少 的比例,尤其缺乏古漢語(上古、中古、 近代等)的語料。這使得古漢語的自然語 言處理一直沒有適切的大型預訓練模型 可用。基於此,我們訓練與發佈了一個專 爲古漢語打造的 BERT 系列模型。我們 的預訓練語言模型與原本的中文 BERT 系列模型相比,能夠成功降低了古漢語 的 perplexity 分數。同時,我們也進一 步開發了不同時代的分詞與詞類標記的 模型,並探究其對於跨時代語料的遷移 學習能力。最後,我們將不同時代模型 的人稱代名詞詞向量 (word embedding) 進行降維, 觀察其不同時代的變異情形。 我們的程式碼發布在https://github.com/ ckiplab/han-transformers °

Abstract

The pre-trained language model has recently dominated most downstream tasks in the NLP area. Particularly, bidirectional Encoder Representations from Transformers (BERT) is the most iconic pre-trained language model among the NLP tasks. Their proposed maskedlanguage modeling (MLM) is an indispensable part of the existing pre-trained language models. Those outperformed models for downstream tasks benefited directly from the large training corpus in the pretraining stage. However, their training corpus for modern traditional Chinese was light. Most of all, the ancient Chinese corpus is still disappearance in the pretraining stage. Therefore, we aim to address this problem by transforming the

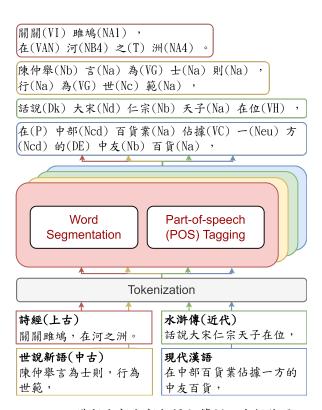


Figure 1: 漢語分詞和詞類標記範例,我們使用四個時代的漢語語料預訓練語言模型,同時應用在分詞和詞類標記的下游任務上。

annotation data of ancient Chinese into BERT style training corpus. Then we propose a pre-trained Oldhan Chinese BERT model for the NLP community. proposed model outperforms the original BERT model by significantly reducing perplexity scores in masked-language modeling (MLM). Also, our fine-tuning models improve F1 scores on word segmentation and part-of-speech tasks. Then we comprehensively study zero-shot cross-eras ability in the BERT model. Finally, we visualize and investigate personal pronouns in the embedding space of ancient Chinese records from four eras. We have released our code at https://github.com/ckiplab/ han-transformers.

關鍵字:漢語、預訓練語言模型、零樣本跨 時代學習

Keywords: Chinese Language Model, Zero-shot Cross-Era Transfer Learning

1 緒論

近年自然語言處理 (Natural Language Processing, NLP) 領域中,大型預訓練的語言模型 (Kenton and Toutanova, 2019; Radford et al., 2019; Raffel et al., 2020) 在許多下游任務中均取得優異的表現。多語言的預訓練模型 (Kenton and Toutanova, 2019; Xue et al., 2021) 也日益出現,能夠支援中文的自然語言處理。常見的中文下游任務包含問答系統 (Shao et al., 2018; Cui et al., 2019)、機器翻譯 (Sennrich et al., 2016)、文本情感分析 (Pontiki et al., 2016; Tan and Zhang, 2008)、文本摘要 (Hu et al., 2015)、詞性標記 (Xue et al., 2005) 和中文分词 (Emerson, 2005; Jin and Chen, 2008) 等等。

得益於多語言的預訓練資料,(Pires et al., 2019) 發現以 104 種語言預訓練的 Multilingual BERT (mBERT) 有 zero-shot 的跨語言 學習能力。以中文問答系統爲例, (Hsu et al., 2019) 實驗得出英文問答資料上做微調的 BERT 模型,測試於中文的問答資料上就能 得到不錯的 F1 值。而同時使用中文和英文的 問答資料做微調,比單獨使用中文的資料表 現有更高的 F1 值。然而多語言的預訓練模型 同時也受限於預訓練階段不同語言文本的資 料量多寡,與多資源的語言相比之下,下游任 務在少資源語言上表現較爲不佳,舉例來説, mBERT (Pires et al., 2019) 的預訓練資料中 最多的是英文 (21%),因此在 XNLI (Conneau et al., 2018) 的任務上英文有最好的表現,相 比之下簡體中文差了大約五個百分點。在現存 大部分預訓練語言模型以英文語料爲主的情況 下,增加中文的預訓練語料來提升中文下游任 務的成績,有著迫切的需求。

簡體中文的領域有已經發布的MacBERT (Cui et al., 2021)使用大量簡體中文語料做為預訓練文本,在各個中文下游任務比原始BERT模型 (Kenton and Toutanova, 2019)有更好的成績,然而在繁體中文的領域中,這部分較為缺乏。特別在古漢語的預訓練語言模型。因此,在這份工作中,我們使用古漢語的語料對BERT系列模型做預訓練,並開發了不同時代的分詞與詞類標記的模型。同時,基於上述多語言預訓練給我們的啓示,我們擬探究中文不同時代的語料訓練出

斷詞和詞性標註 (撷取自堯典)

乃 (DL) 命 (VF) 義 (NB1) 和 (NB1) , 欽若 (VH1N) 昊天 (NA1) ; 歴 (DA) 象 (VP) 日月 (NB2) 星辰 (NA2) , 敬 (DV) 授 (VD) 人時 (NA5) 。

Table 1: 漢語語料庫標註資料

來的模型是否具有遷移能力,以及綜合不同時 代的語料一起訓練是否會有助於各時代的下游 任務表現等問題。

針對上述需求與問題,我們使用中央研究院 的四個時代的漢語文本語料庫作爲預訓練的 資料,並使用其對應的標記語料做爲下遊任 務的調適訓練和測試,包含上古漢語1(先秦到 西漢)、中古漢語2(東漢魏晉南北朝)、近代漢 語3(唐五代以後) 和現代漢語平衡語料庫4。各 時代標記語料均含有人爲標記的斷詞和詞類, 如表1所示,第一行的「乃」是關聯副詞、「命」 是複雜雙賓動詞 (外動複),而第二行開頭的 「欽若」是兩個中文字組成的狀態不及物動詞 (内靜),我們使用這些標記資料來做中文的分 詞與詞類標記任務。透過這四個時代的 BERT 預訓練與微調訓練,我們擬探討不同時代語料 能否互相增進學習表現。實驗發現,對於分詞 任務,不同時代語料的確能夠互相增進學習, 證明模型的遷移能力,而對於詞類標記任務, 由於詞類標記在四個時代的標記分類不盡相 同,像是上古語料中有 382 種 (89%) 詞類標 記並未在其他三個時代語料出現,因此我們發 現詞類標記任務就不具備這樣的遷移能力。

之後的章節中,第2章將介紹如何將中央研究院的漢語標記語料庫轉換成 BERT 模型 (Kenton and Toutanova, 2019) 的資料型態,我們撰寫程式轉換三種任務對應的不同輸型的訓練方法,使用我們資料集預訓練中文語言模型,以及分別微調 (Fine-Tune) 預訓練模型在兩個下游任務模型。接著我們在第4章節給果,除了三種模型在不同時代資料集的可能是不同時代資料集的可能與型在不同時代資料集的可以與試結果和討論,我們也對不同時代的詞嵌代詞 (personal pronouns) 在語言模型的詞嵌行觀察與分析。本論文主要有以下四種貢獻:

 $^{^{1}\}rm http://lingcorpus.iis.sinica.edu.tw/cgibin/kiwi/akiwi/kiwi.sh$

²http://lingcorpus.iis.sinica.edu.tw/cgibin/kiwi/dkiwi/kiwi.sh

³http://lingcorpus.iis.sinica.edu.tw/cgibin/kiwi/pkiwi/kiwi.sh

⁴http://asbc.iis.sinica.edu.tw/

詞語	詞類	頻率	Vocab. set
_	S	2108	-
-	VH1	23	## 夫
_	VP	47	## 旦
一夫	NA1	2	## 再
一旦	NA5	3	## 氣

Table 2: 轉換標記資料成 BERT 字彙表

Domain	#Lexicons	Size of Vocab.
上古	41595	8781
中古	25956	7448
近代	36906	8660
現代	144655	12563

Table 3: 四個時代字彙統計

- 使用四個時代的漢語語料來預訓練語言模型,其中使用古漢語語料是自然語言處理領域的首次嘗試。我們預訓練的語言模型符合 HuggingFace 平台上 transformers 套件的使用格式,進一步促進繁體中文語言處理的研究發展。
- 微調預訓練語言模型在下游任務,包含中 文分詞與詞類標記任務,古漢語的自動化 分詞與標記能幫助歷史語言方面的研究。
- 啓發於 (Pires et al., 2019; Hsu et al., 2019) 發現語言模型在下游任務有 zero-shot 的 跨語言學習能力。我們想觀察語言模型在 我們的兩個下游任務上是否有 zero-shot 跨時代漢語學習能力。實驗發現在中古語 料數量最少的情況下,拿訓練在近代語料 上的分詞模型直接測試在中古語料上,和 使用中古語料訓練的分詞模型相比達到可 競爭的 F1 分數,發現模型具有 zero-shot 跨時代漢語學習能力。
- 視覺化語言模型中人稱代詞的詞嵌入向量,觀察漢語人稱代詞裡詞跟詞 (例如「濃」和「吾」),在四個時代詞空間裡相對位置的一致性。

2 資料處理

2.1 建立模型字彙表

我們撰寫程式將漢語語料庫提供的詞語表轉成 BERT 格式的字彙表 (如表2),這些在字彙表 中的字詞 (token) 可以重組成表2左邊的所有 詞語。BERT 的字彙表使用 Google NMT (Wu et al., 2016) 提出的 WordPiece 斷詞方法,將 原來的 words 拆成更小微度的 wordpieces,可

Original annotation							
ーロ若 (T) 稽 (VC2) 古 (NA5) :							
帝堯 (NB1) ロ (VG) 放勳 (NB1) 。							
斷詞	日	若	稽	古	:		
分詞	В	I	В	В	В		
詞類	${ m T}$	${ m T}$	VC2	NA5	-		
斷詞	帝	堯	日	放	勳	0	
分詞	В	I	В	В	I	В	
詞類	NB1	NB1	VG	NB1	NB1	-	

Table 4: 斷詞和詞性標註訓練範例

以有效處理不在字典裡頭的詞語。而中文方面是字符單位 (character-level) 的斷詞,有 ##前綴的字詞即爲 wordpieces。如下表2的「一夫」,長度爲 2 個字符,由「一」和「## 夫」兩個字詞組成。

轉換完成的數量集大小如表3所示,原本上古中有 41595 個詞語 (Lexicon),我們將之轉換縮減成總共 8781 個字詞的 BERT 字彙表,也就是這 8781 個字詞即可表示左邊的 41595種詞語。中古、近代和現代分別將各自的詞語從 25956、36906 和 144655 透過我們的程式轉成 7448、8660 和 12563 個字詞。四個資料集合計有 37452 個字詞,這些字詞可以完整表示原本四個時代漢語的所有詞語。

2.2 建立分詞和詞類標記訓練資料

我們撰寫程式將漢語語料庫提供的分詞和詞類標記資料轉換成模型訓練資料,如表4所示。最上面是語料庫提供的人工標記資料,下面是我們轉換完成的分詞任務和詞類標記任務的於對之,分詞模型的標記轉成以BIO格式來表示。「曰若稽古:」是一句話,而其內方,所以分詞模型輸入「曰若稽古:」時,我們看望分詞模型輸出「BIBBB」。而詞類標記模型輸出「BIBBB」。而沒與一類標記模型輸出「TTVC2NA5」。表示「母若」(T)是語助詞,「稽」(VC2)是準動作單數詞(準外動),「古」(Na5)是時間詞。

2.3 新增古漢語字彙到實驗模型

我們將轉換完的古漢語字彙表新增到實驗模型,擴充原模型的中文詞彙表,實驗模型裡新增的四個時代詞彙字詞個數如表5所示。我們的實驗總共實作和比較在五種模型上,其中三個 ckiplab/開頭的模型 (表5上方) 是中文詞庫小組5之前只訓練在現代的語料上 (ZhWiki

⁵https://github.com/ckiplab/ckip-transformers

Model		中古	近代	現代
ckiplab/albert-tiny-chinese	2673	1041	941	2635
ckiplab/albert-base-chinese	2673	1041	941	2635
ckiplab/bert-base-chinese	2673	1041	941	2635
roberta-base (Liu et al., 2019)	8781	7444	8660	12563
bert-base-uncased (Kenton and Toutanova, 2019)	8371	6983	8212	12134

Table 5: 實驗模型擴充後的字彙數量

輸入句子	則必有穿窬拊楗、抽箕踰備之姦;
制入りつ	爲孔子之窮於陳、蔡而廢六藝,
	'則', '必', '有', '穿', '[UNK]', '[UNK]', '[UNK]', '丶', '抽', '箕', '[UNK]', '備',
新增字詞前	'之', '姦', ';' (from ckiplab/bert-base-chinese)
利石丁时的	'[UNK]', '[UNK]', '子', '之', '[UNK]', '[UNK]', '陳', '丶', '[UNK]', '[UNK]',
	'[UNK]', '[UNK]', '[UNK]', ',' (from bert-base-uncased)
新增字詞後	'則','必','有','穿','[UNK]','拊','楗','丶','抽','箕','踰','備','之','姦',';'
(我們模型)	'爲', '孔', '子', '之', '窮', '於', '陳', '丶', '蔡', '而', '廢', '六', '藝', ','

Table 6: 新增字詞 (tokens) 前後的模型斷詞結果

與 CNA 資料集)。另外我們挑選兩個 BERT 的系列模型 (表5下方),可以發現 bert-base-uncased (Kenton and Toutanova, 2019) 模型在作者的原始訓練資料中已含有部分中文字彙表,而 roberta (Liu et al., 2019) 的模型完全沒看過中文資料,因此我們轉換完成的字彙表被完全擴充到 roberta 的模型字彙表中。

新增古漢語字詞到模型後,輸入的文句會先 經由模型的 tokenizer 做斷詞處理,在 BERT 模型中不認識的字詞會以「[UNK]」表示。如 表6所示,當輸入「則必有穿窬拊楗、抽箕踰備 之姦;」到 ckiplab/bert-base-chinese 模型時, 因爲新增字詞前模型不認識「窬」、「拊」、「楗」 和「踰」,所以斷詞結果皆顯示爲「[UNK]」, 在語言模型内映射這四個字到向量空間時,皆 以「[UNK]」對應的詞嵌入向量表示,導致模 型理解混亂。而在我們新增從漢語語料庫提供 的 Lexicon 轉成的字詞到模型後,表6最下列 在這個範例中模型多認識了「拊」、「楗」和 「踰」三個字。而輸入「爲孔子之窮於陳、蔡 而廢六藝 ,到 bert-base-uncased 模型時,可 以發現這輸入範例中原始模型認識的中文字非 常少,只有「子」、「之」、「陳」以及標點符 號 (表6中間),在我們新增字詞到模型後,此 範例所有的中文字都成功斷詞,在之後模型裡 能以各字詞獨有的詞嵌入向量表示。

3 方法

3.1 中文語言模型

我們採用 (Kenton and Toutanova, 2019) 預訓練語言模型的方法,稱爲 Masked LM

(Masked Language Modeling) 的訓練技巧,目的是訓練詞嵌入向量來學習雙向的語言資訊。以表6舉例,當我們輸入文句「曰若稽古:帝堯曰放勳。」時,在 Masked LM 中,在斷詞後的 11 個字詞中,每個字詞有 15% 的機率被替換成「[MASK]」,而模型的目標是預測出被遮起來的原字詞是什麼。在這個任務目標上,模型必須得理解「[MASK]」前後兩端的語言資訊才能正確推論出原本被我們替換成「[MASK]」的字詞原本是什麼,因此可以訓練出有雙向資訊的上下文詞嵌入向量。

依照原始 BERT(Kenton and Toutanova, 2019) 的說法,如果將所有選定的 token 皆替換成「[MASK]」,會導致模型學到只需要在遇到「[MASK]」時才做預測,其他的字詞皆視作上下文來訓練。因此實作上,作者設定的是 15% 被選定替換的 tokens 中,80% 替換成「[MASK]」、10% 是字彙表中隨機的 token 和 10% 保留原本的 token。我們也依循此設定進行古漢語文字的 Masked LM 訓練。

我們將四個時代的資料以 10:1:1 的比例切割為訓練資料集 (train)、開發資料集 (dev)和測試資料集 (test set),各資料集句子數量如表7所示。在訓練階段時使用訓練資料集來做訓練,每 500 個訓練步數使用開發資料集做測試並將表現最好的模型儲存下來。訓練完成後,我們統一測試於各時代測試集上作爲實驗結果。表7中最下方的 merge,是我們合併四個時代的資料集來訓練模型,並分別測試在四個時代的測試資料集上。

Domain	Train	Develop	Test
上古	460170	43890	43060
中古	262395	22688	31277
近代	698749	61934	76106
現代	1175516	109715	110718
Merge	2596830	238227	-

Table 7: 語言模型訓練資料

Dataset	Train	Develop	Test
上古	3274921	326074	325511
中古	2169011	214735	220383
近代	6511768	645702	659317
現代	15824246	1576439	-

Table 8: 分詞標記模型訓練資料

3.2 中文分詞模型

當模型輸入表9的斷詞結果(以中文字符爲單位),分詞模型會對每個字詞輸出 IBO 格式的標記結果,如表9中間所示。分詞訓練的任務上,對每一個字詞做分類,區分是應的人或。最後經過後處理將 BI 標記對最高之字串轉成我們可以得到分詞結果,如表9最下行人,」後我們可以得到分下條一次,「所以,「所不為」、「所以,「所不為」、「所以,「所不為」、「所不為」、「所不為」、「所不為」、「所以得到」、「於」、「所以,「所不為」、「所以,「所不為」、「所以,「所以,」、「所以,「所以,」、「所以,」、「所以,「所以,」、「所以,」、「所以,」、「所以,」、「所以,」、「所以,」、「所以,」、「所以,」、「以為與3.1語言模型訓練相同),之後每一筆資料以詞語作爲單位。

3.3 詞類標記模型

詞類標記任務和分詞任務相同,同樣可視爲 以字詞爲單位的分類問題。但不同於分詞模 型只有 B、I和 () 共三種類別,四個時代語 料中總共包含 518 種詞類類別。上古語料使 用當中 428 種詞類爲最多,因爲在漢語語料 庫人工標記時,對上古語料的動詞後有加註 字母來表示語境和特徵。而中古、近代和現代 資料集則分別使用 98、98、76 種詞語標記。 圖2表示他們標記的分布情形。詞類標記任務 如表9所示,當輸入句子「朱儒問徑天高於脩 人,」後模型會輸出每個字的詞類標記,幫 助我們理解句中每個詞語的作用。如範例所 示:「朱儒」(NA1) 是有生名詞,「問」(VE) 是後接句賓或動詞組的動作及物動詞 (外動 子),「徑」(U) 是待分析詞句,「天高」(NI) 是抽象名詞及衍生名詞,「於」(P)是介詞, 最後「脩人」(NA1) 是有生名詞。得知古漢語 中一句話裡每個詞的作用後,能更方便我們理

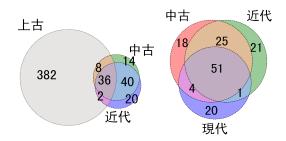


Figure 2: 各時代語料中詞類標記集

解此句含意。模型訓練資料統計結果如表10, 以句爲單位使用之前語言模型訓練同樣的資料 集切割,每一筆訓練資料如表4所示。

4 實驗模型

所有實驗模型的大小皆使用 (Kenton and Toutanova, 2019) 中定義,例如 BERT-base 就是總共 12 層、768 維的 hidden vector、12 個 self attention head 等基於 transformer 架構的模型。我們實驗總共包含三種類別 (baseline, our indiv., our merge) 的模型,以訓練資料作爲區分方式,訓練完成後皆分開測試於四個時代的測試集上來進行比較。

4.1 之前發表模型 (baseline)

在預訓練語言模型、分詞和詞類標記等三個任務上,直接取用原發表模型的參數初始化任務模型,不進行任何訓練且凍結此任務模型的參數。經過模型的 inference 模式,在四個時代的測試資料集上進行測試,此結果作爲我們的baseline 模型,跟後面兩種經由我們訓練後的模型結果作比較。

4.2 訓練於獨立時代語料模型 (our indiv.)

首先,在模型初始化上,我們的預訓練語言模型使用已發佈的 BERT 的參數初始化,而分詞和詞類標記等下游任務則用我們預訓練好的 bert-base-han-chinese 模型初始化。接著進行微調 (Fine-Tune),分別經由我們四個時代的資料集做訓練,同時在訓練當中透過同時代的開發資料集選出表現最好的模型 checkpoint。在這個方法中,每個時代都會訓練出各自的模型,以我們的 bert-base-han-chinese舉例 (表11中間),對應上古、中古、近代和現代都有一個各自微調的 bert-base-han-chinese模型,總共有四個模型,各自測試其對應時代的測試集。

4.3 訓練於合併時代語料模型 (our merge) 模型初始化方式同4.2,不同的是訓練時使用 的資料集,三個任務的訓練和開發資料集分

輸入句子				朱儒	問徑	天高石	於脩ノ	ζ,			
斷詞結果	朱	儒	問	徑	天	高	於	脩	人	,	
詞類模型輸出	NA1	NA1	VE	U	NI	NI	Ρ	NA1	NA1	-	
分詞模型輸出	В	I	В	В	В	I	В	В	I	В	
後處理結果	朱儒	(NA1)	問 (V)	王) 徑	(U)	天高	(NI)	於 (P)	脩人	(NA1)	,

Table 9: 分詞標記和詞類標記範例

Dataset	Train	Develop	Test
上古	459910	43859	42877
中古	262271	22668	31260
近代	698746	61933	76106
現代	1175516	109715	110718
Merge	2596443	238175	-

Table 10: 詞類標記模型訓練資料

別使用表7、8和10最下方的合併 (merge) 資料集。不同於獨立時代語料的模型,我們的bert-base-han-chinese 只有一個模型。

5 實驗結果與討論

這章節是我們的實驗結果,在語言模型上,含有在混淆度上測試的結果和視覺化人稱代詞的分析。之後對兩個下游任務上有 F1 分數的結果和 zero-shot 跨時代學習能力的討論。

5.1 語言模型

語言模型在各時代語料測試集上的結果如表11所示。粗體是各時代語料上最好的結果。最上方是 baseline 模型,包括 google 的BERT 模型和來自 ckiplab 發布的繁體中文語言模型。ckiplab 的三個 BERT 系列模型只訓練在現代的語料上 (ZhWiki 與 CNA 資料集),因此在現代測試集上表現不錯,但在上古、中古和近代測試集上都明顯表現不佳。

表11中間是4.2提到的獨立時代語料訓練結果,與 ckiplab 的模型相比,在同樣訓練於現代語料上,表現最好的 bert-base 模型降低4.23 的混淆度 (8.85 vs 4.61)。而在上古、中古和近代三個時代測試集上跟 baseline 相比皆有顯著的進步,例如上古原本 bert-base 模型的混淆度高達 233.64,經過上古資料的訓練後,我們的 bert-base 模型能降到 24.76。

最下方是使用我們合併的四個時代資料集進行訓練,可發現整體趨勢上,除了albert-tiny模型之外,merge資料集訓練的模型表現較佳,雖然以bert-base來看,測試於上古和近代輸給各自時代訓練的模型,但差距並不大(上古差6.42、近代差2.24),在中古有9.39的進步,在資料集數量大的現代語料上

也有 0.11 的些微進步,而在 roberta-base 模型上,上古和中古皆有進步,但近代和現代略輸 0.36 和 1.44,但整體結果上 merge 資料的確對語言模型訓練有一定的幫助,在中古資料集數量最少的情況下 (表7),在中古測試集上 merge 的好處尤爲明顯。在後續的實驗中,雖然整體是 roberta-base 模型較好,但考慮到bert-base 模型與其相差不大,而且是唯一全部贏過 baseline 的模型 (roberta-base 在現代上比 baseline 差),因此在下游任務初始化我們皆選用 merge 的 bert-base-han-chinese。

語言模型交叉測試在不同時代語料上,如表12所示,可以發現在同樣模型 (皆為 bertbase)下,整體來看 merge 的結果最好,四個時代皆贏過我們的 baseline 模型。與獨立時代語料訓練相比,merge 模型另一個優點是,使用上不需背景知識先判斷資料隸屬於哪個時代來使用對應的時代模型。

5.2 向量空間中的人稱代詞

我們對訓練好的語言模型詞向量空間做分析,挑選人稱代詞做 PCA 和 t-SNE 降維,結果如圖3所示。當中包含第一人稱的「余、我、吾」,第二人稱的「你、汝、儂」,第三人稱的「你、汝、儂」,第三人稱的「你、汝、儂」,第三人稱的「你、汝、儂」,第三人稱的「你、汝、儂」,第三人稱的「你、汝、儂」,第三人稱的「你、汝、儂」,第三人稱的「你以禮學」的自己,以為其也。圖3左邊 PCA 降維可以看到同一個詞在不同時代的模型距離相近」,表示同樣的詞在不同時代中相對於其他問題不同樣的詞。圖3右邊經由 t-SNE 降相的結果,可以觀察到類似同心圓的分布,最中間最聚集的是 google/bert-base-chinese 模型,由內往外分別是上古 (橋)、現代 (紫)、中古(錄)和近代 (紅)。

5.3 分詞模型

分詞標記的實驗結果如表13所示。表13最上方是中文詞庫小組之前發布的中文分詞模型,作爲我們的 baseline,在現代測試集上有最好的表現,但在其他三個時代語料上卻表現很差。我們的分詞模型使用5.1預訓練完成的 our/bert-base-han-chinese(merge) 初始化

T	TD	$\textbf{Test-set, Perplexity} \downarrow$				
Language Model	Train-set	上古	中古	近代	現代	
google/bert-base-chinese		167.7257	268.6131	187.615	10.5801	
ckiplab/albert-tiny-chinese		627.9473	780.1218	563.2704	34.9042	
ckiplab/albert-base-chinese	-	359.8375	520.2965	388.5624	39.7566	
ckiplab/bert-base-chinese		233.6394	405.9008	278.7069	8.8521	
our/albert-tiny-han-chinese	our indiv.	48.0267	139.6061	79.7592	13.0466	
our/albert-base-han-chinese	our indiv. 上古/中古/近代/	37.0239	112.6525	65.0357	7.4371	
our/bert-base-han-chinese	現代	24.7588	70.6244	46.8308	4.6143	
our/roberta-base-han-chinese	<i>7</i> 010	20.9822	64.5587	30.6159	12.3762	
our/albert-tiny-han-chinese		68.7009	107.9195	93.0871	14.2858	
our/albert-base-han-chinese		50.7889	87.0224	74.4756	7.6285	
our/bert-base-han-chinese	our merge	31.1807	61.2381	49.0672	4.5017	
our/roberta-base-han-chinese		20.6797	37.5194	30.9787	13.8190	

Table 11: 預訓練語言模型在各時代測試集上的混淆度分數,數字越小越好。最上方是我們的 baseline 模型,挑選原始 google 發布的 bert-base-chinese 模型和三個來自中文詞庫小組發布的預訓練語言模型,不進行訓練 (zero-shot) 測試在四個時代語料上。中間部份是我們使用獨立時代語料訓練完成的語言模型,模型訓練集和測試集皆來自同一時代語料。最下方是我們合併四個時代語料訓練的語言模型,統一使用四個時代合併的語料進行訓練,再分別測試於四個時代測試集的結果。

T M - d - l	The :	$\textbf{Test-set, Perplexity} \downarrow$				
Language Model	Train-set	上古	中古	近代	現代	
ckiplab/bert-base-chinese	-	233.6394	405.9008	278.7069	8.8521	
	上古	24.7588	87.8176	297.1111	60.3993	
	中古	67.861	70.6244	133.0536	23.0125	
our/bert-base-han-chinese	近代	69.1364	77.4154	46.8308	20.4289	
	現代	118.8596	163.6896	146.5959	4.6143	
	merge	31.1807	61.2381	49.0672	4.5017	

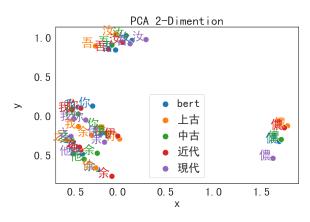
Table 12: 交叉測試同架構語言模型在不同時代測試集上的混淆度分數

TX7C T / 1 . 1	(T)	$\overline{\text{Test-set, F1}^{\uparrow}}$				
WS Model	Train-set	上古	中古	近代	現代	
ckiplab/bert-base-chinese-ws	-	86.5698	82.9115	84.3213	98.1325	
	上古	97.6090	88.5734	83.2877	70.3772	
	中古	92.6402	92.6538	89.4803	78.3827	
our/bert-base-han-chinese-ws	近代	90.8651	92.1861	94.6495	81.2143	
·	現代	87.0234	83.5810	84.9370	96.9446	
	merge	97.4537	91.9990	94.0970	96.7314	

Table 13: 交叉測試中文分詞模型在不同時代語料上的 F1 分數

DOC M. 1.1	Train-set	Test-set, F1↑			
POS Model		上古	中古	近代	現代
	上古	91.2945	-	-	-
	中古	7.3662	80.4896	11.3371	10.2577
our/bert-base-han-chinese-pos	近代	6.4794	14.3653	88.6580	0.5316
	現代	11.9895	11.0775	0.4033	93.2813
	merge	88.8772	42.4369	86.9093	92.9012

Table 14: 交叉測試詞類標記模型在不同時代與語料上的 F1 分數



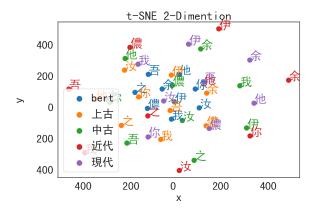


Figure 3: 左圖、右圖分別是對不同人稱代詞的 embedding 做 PCA、t-SNE 降維的視覺化,測試於五種模型,包含原始 bert-base-chinese 與我們分別使用四個時代資料集訓練的語言模型。

Number of POS types						
上古	中古	近代	現代	merge		
428	98	98	76	518		

Table 15: 各資料集的詞類標記數量

5.4 詞類標記模型

詞類標記模型的表現如表14所示,可以發現如 同分詞模型的結果 (表13),單獨時代語料訓 練的模型在同時代語料上有最好的表現。但 與5.3不同的是,獨立時代語料訓練完的模型 並沒有跨時代學習的能力,我們推測原因是: 詞類標記任務與分詞任務雖然都是分類問題, 但分詞任務的輸出只有三個類別 (BIO),且四 個時代皆是同樣的;但詞類標記任務在表10和 圖2可以發現各時代語料的類別集合皆不完全 相同,尤其在上古的部分,所有 428 個類別 裡有高達 382 個 (89%) 的類別是它自己獨有 的,因此訓練在上古的詞類標記模型完全無法 應用在其他三個時代。而其他三個時代單獨訓 練的模型測試在不同時代測試集上表現也大幅 下降。相對來看,則可以發現 merge 的模型 整體來看較爲穩定,雖然受限於資料數量不平

衡的原因 (表10),在中古數量最少的情況下, 導致 merge 的模型測試在中古上表現較差。

5.5 零樣本跨時代學習能力

綜合表7、8和10來看,可以發現訓練數量是現代最多、近代次之,而中古最少。在這樣的環境下,語言模型的結果 (表12) 可以看到只納練在近代語料的模型,直接應用在中古語料上 (77.41),比訓練在中古且測試在中古語料上 (77.41),比訓練在中古且測試在中古語料 的結果 (70.62) 還要更好,而使用 merge 資的練的結果 (表13) 也可去語料上則有最好的模型測試在中古語料上也有可說等在近代的模型測試在中古語料上也有可說等在近代的模型測試在中古語料上也有可說等在近代學習的能力。唯一例外的是調標標準,受限於不同時代的類別相差過大,因此無法成功跨時代語料學習。

6 結論

現今已發布的預訓練語言模型裡,大多數都是 基於英文的模型,在其他語言中建構強大預訓 練語言模型的嘗試較少。在古漢語這塊,更是 完全沒有可使用的模型。因此我們考量到自然 語言處理領域上,目前繁體中文的缺乏和古漢 語的缺失。我們使用包含四個不同時代的漢語 語料庫來預訓練語言模型,同時成功應用在分 詞和詞類標記的下游任務上。除此之外,Bert 架構模型除了之前研究發現的零樣本跨語言學 習能力外,在我們的實驗中發現語言模型在掩 碼詞 (masked word) 預測和分詞任務上都具有 零樣本跨時代學習能力。最後我們使用 PCA 和 t-SNE 對四個時代的人稱代詞詞向量降維 並視覺化呈現,發現不同時代模型中詞跟詞之 間有相同的相對位置關係。在未來的研究上, 不同於 Bert 是基於 encoder 的模型, 我們可 以嘗試使用現存的其他基於文字生成的模型。

Acknowledgments

本研究由中央研究院數位人文研究計畫 (2379-111237j) 贊助支持。古漢語標記語料由中央研究院語言所提供,特此感謝。感謝國研院國網中心提供計算與儲存資源,協助本研究順利進行。

References

- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2475–2485.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.
- Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5883–5889.
- Thomas Emerson. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*.
- Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5933–5940.
- Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972.
- Guangjin Jin and Xiao Chen. 2008. The fourth international chinese language processing bake-off: Chinese word segmentation, named entity recognition and chinese pos tagging. In *Proceedings of the sixth SIGHAN workshop on Chinese language processing*.

- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pretraining of deep bidirectional transformers for language understanding. In *Proceedings of* naacL-HLT, pages 4171–4186.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001.
- Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings* of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 371–376, Berlin, Germany. Association for Computational Linguistics.
- Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. Drcd: a chinese machine reading comprehension dataset. arXiv preprint arXiv:1806.00920.
- Songbo Tan and Jin Zhang. 2008. An empirical study of sentiment analysis for chinese documents. Expert Systems with applications, 34(4):2622–2629.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.