

基於少量中文文本的新短語發現與關鍵詞抽取

張郁
資工碩一

林順天
資工四 A

何冠緯
資工四 B

楊佳峻
資工三 B

國立中央大學 資訊工程學系

Department of Computer Science and Information Engineering, National Central University

Abstract

新詞發現是中文自然語言處理的基礎任務之一。該任務的目標是從語料中抽取較常搭配出現的詞語或短語，組成詞典，以便下游的分詞、關鍵詞抽取、分類等任務使用。既有的新詞發現方法主要用於抽取 2-3 個字的中文詞語，並且需大量文本才能獲得較好效果。本研究希望突破這兩項限制，利用預訓練語言模型，在 100 篇文本中進行微調，結合傳統統計方法進行短語發現、篩選及關鍵詞抽取。結果，我們的模型在短語篩選中達到了 93.5 的 F1 分數。由新短語組成的詞典也可以提升關鍵詞抽取效果。我們的原始碼公布在：<https://github.com/fireindark707/WIMU2022-new-keyphrase-discovery-extraction>

1 介紹

基於詞袋模型 (Bag of words model, BoW) 的技術是自然語言處理領域的傳統方法，雖然近年來基於深度學習、預訓練語言模型的方法在大部分任務上的表現都超過了詞袋模型。但是深度學習的缺點——需要大量運算資源、推論速度較慢、不夠穩健——也是明顯的。因此，在實務應用中，詞袋模型仍然在被大量使用。而建構詞典則是詞袋模型的方法基礎，越適應領域、越完整的詞典越可以讓詞袋模型在下游的分詞、關鍵詞抽取、分類、分群等任務中表現得更好。

本研究就是以在少量中文文本上建構領域詞典作為目標。並且，我們希望超越傳統方法以 2-3 字的詞語為主的詞典建構方法，試圖建構 3 字-8 字長度的短語詞典。這樣生成的詞典更容易包含一些領域的專有名詞、專業詞彙或實體名稱，並且因為長度較長也會包含更多內容。此外，結合了新的短語詞典的方法在關鍵詞抽取等下游任務中可能有更好的表現。

雖然較長的短語詞典可能有較好的效果，但是長詞通常在語料中的出現頻率較低，因此傳統依賴詞頻 (Term frequency, TF) 進行統計的方法無法獲得較好的結果。我們提出的解決方案是：對傳統統計方法設定較低門檻，給出大

量品質不一的候選短語，並結合預訓練語言模型，對候選短語進行篩選、判斷，從而生生成規模較大、質量較高的短語詞典。

我們的主要貢獻如下：

- 在少量（100 篇左右）文本的基礎上完成新短語發現。
- 結合預訓練語言模型提升短語詞典品質。
- 成功抽取 3-8 字的較長短語並判別短語領域。

2 文獻回顧

2.1 基於統計的新詞發現

依任務的不同，分詞精度可能會受現有分詞工具所限，Statistic method 常用於建立一個可能為新短語的候選詞集，並透過模型分析後發現新詞，亦可直接用來生成新語料庫，或是先創建候選詞集再過濾至同義詞集，最後生成類別相近的複合單詞、專業單詞 (Zhang et al., 2019)。再進階的則是先分別計算出五個特徵，並根據人工標註結果使用邏輯回歸計算特徵權重 (Chen et al., 2017)。

常見評估方法有 Left and right neighbor information entropy (branch entropy)。TF-IDF、Mutual Information、Word2Vec 模型，分別用以劃定斷詞邊界、計算詞頻、分析相關度、計算相似度，有許多論文將上述方法組合運用，計算自己的相似度，並設閾值過濾字詞以生成新詞庫，其他還有些類似的工具，例如：計算詞頻的 BM25 (Correia et al., 2018) 等...

在中文新詞發現研究方法中曾提出過基於新詞特徵的相似度判斷原則，並結合 Mutual Information 和相似度引入相似度增強互信息的概念 (Shang, 2019)，雖然該方法對於繁體中文的處理和特殊符號的過濾不完善，使其由停用詞組成的詞串並不都是非法串，但確實提供了一種有效的過濾方法供我們參考。

在中醫藥領域，文本語料庫有一些現代文文與文言文的路徑交叉的情況，分詞缺乏人工標記的語料庫也很難獲得，因此很難使用監

督方法來訓練應用於中醫的分詞器。為了解決此問題, Jia 等人使用了一種無監督方法, 以 Entropy(熵) 為特點, 結合中醫領域詞典, 構建了一個中醫文本特定的分詞器 (Jia et al., 2019)。

2.2 基於機器學習的新詞發現

雖然基於統計的新詞發現/切詞方法已經成功運用於下游任務, 但中文的語言複雜性以及 OOV (Out-of-vocabulary) 問題讓既有方法的效果不盡如人意。不少研究者開始運用機器學習的方法, 例如 CRF 及神經網絡模型進行新詞發現, 例如在文本序列中取得較好效果的 LSTM 等 RNN 系列模型, 另外, Xu 等人集成了 GRNN 和 LSTM 以進行更深入的特徵提取 (Xu and Sun, 2016)。

Jie Yang 等人使用一個擴展的 LSTM 版本 Lattice LSTM 進行分詞, 該方法可以將模糊的詞語資訊融合進以字為基礎的 LSTM 模型, 在標準資料集上達到當時最好的成績 (Yang et al., 2018)。後續, Huang Kaiyu 等人使用 Word-Lattice-Based CRF、Marginal Probability Strategy 等非監督方法進行無監督的新詞發現, 效果超過統計方法及 LSTM 為基礎的模型 (Huang et al., 2021)。這些方法都考慮到了中文分詞的模糊性, 並在加入全文本資訊等額外特徵後提升了效果。

2.3 基於預訓練語言模型的無監督詞庫構建

2018 年, BERT 等預訓練語言模型陸續發表, 相較於傳統的 Word Embedding, 預訓練語言模型可以更好地獲得上下文語境下的詞語表示。因此, 在大部分 NLP 任務中, 基於預訓練語言模型的方法是目前的 SOTA。同時, 也有學者發現, 因為預訓練語言模型在預訓練階段讀取了大量文本, 其本身具有一些類似知識圖譜的功能。我們可以據此推測, 預訓練語言模型也適合用來進行無監督的詞庫構建。

這部分相關研究還不多, 但已有一些發展。例如, 生醫領域的術語的近義、同義關係對於文本探勘非常重要, Elliot Schumacher 等學者運用了 ELMo 和 BERT 獲得詞語的表示後進行無監督的醫學同義詞發現 (Huang et al., 2021)。此外, 生醫領域中還有一些重要的術語需要被識別, 但其出現次數一般較少, 難以透過傳統的方法抽取。Oskar Jerdhaf 等學者使用 BERT 對電子病歷中的人工植入物相關術語進行詞嵌入表示, 並成功地將其聚類進行後續分析 (Jerdhaf et al., 2021)。這類任務被稱為 Focused Terminology Extraction。

Lee 等人在自行挖掘的生醫語料庫引入 BERT 進行預訓練, 並運用 WordPiece Tokenization, 解決專有領域之詞條無法在詞庫表查找

的新詞 (Out-Of Vocabulary, OOV) 問題, 後續實驗也在生物醫學的下游任務, 例如: 命名實體識別、關係抽取及問答系統等方面有良好表現 (Lee et al., 2020)。

2.4 關鍵詞抽取

關鍵字擷取是將段落中常出現或是可代表的字詞擷取出來。最基礎的方式像是監督式學習, 雖然精確度高, 不過由於需要常常更新自建辭典在人力上耗費甚大。而非監督式學習則是雖然精確度較低, 不過就不需要標記的工作, 更多的是算法與統計。

常見的非監督式學習有 TextRank (Mihalcea and Tarau, 2004)、YAKE (Campos et al., 2020)、RAKE (Rose et al., 2010) 等等。而近期成效較好的監督式學習則有 KeyBERT (Grootendorst, 2020), 結合了 BERT 與 cosine similarity, 讓找到的關鍵字更貼合文章。

3 資料集

因為沒有搜尋到該問題的公開資料集, 我們將使用一個自己收集的少量文本資料集, 其中包括約 200 篇與中國勞工、勞動相關的新聞與評論文章。來源包括微信、搜狐新聞、澎湃新聞、網易新聞等大型網站, 平均每篇文章約 4000 字符長度。我們從該資料集中抽取 100 篇當作訓練資料集, 其餘作為測試資料集。

資料將會以長度為 3 字以上的短語為單位進行標註, 標註項包括 (1) 是否是一個合適的短語 (2) 是否與勞工勞動有關 (3) 是否與一實體相關, 例如公司名、人名、地名。標註結果共四個類別, 分別為 O、General、Labor、Name。O 代表該詞非合適的短語, General 代表該詞為與勞工勞動無關也非實體相關的其他短語, Labor 代表該詞是一個與勞工勞動相關的短語, Name 代表該詞是一個實體短語。例如, 「畢業生就業」會被標示為 Labor、「高質量發展」會被標示為 General、「李向倫」被標示為 Name。O 通常有以下三種情況: (1) 由兩個及以上分別的詞組成, 且其組成短語更適合拆分理解, 例如「接受採訪」、「司機們」、「疫情嚴重」(2) 短語及狀態表示, 例如「按最低工資」(3) 本身為合適的短語, 但僅為狀態、語氣表達, 與現實概念較缺乏連結, 例如「记忆中」、「一起去」、「感受到」。

候選短語預計由 4.1 所述的統計方法得到, 我們共在訓練文本中獲得 692 個候選短語、測試文本中獲得 822 個候選短語。Table 1 展示了訓練文本中所抽取的候選短語的在長度和標註類別兩個方面的詳細分佈。

Type	Length						Total
	3	4	5	6	7	8	
O	209	93	5	2	0	0	309
General	48	129	24	4	1	1	207
Labor	16	86	22	8	1	0	133
Name	29	10	2	1	1	0	43
Total	302	318	53	15	3	1	692

Table 1: 訓練文本候選短語的長度與類別分佈

	tf	agg_coef	max_E	min_E
threshold	5	60	1.5	0.5
非政府組織	39	1839.35	4.086	3.453
公眾號	42	1821.91	3.237	2.612
永壽路	5	6791.46	1.921	1.370
重金屬汙染	17	4551.58	3.499	2.651
新冠病毒	9	1030.44	2.725	2.641
超齡農民工	5	298.76	1.521	1.370

Table 2: 訓練文本閾值與候選詞示例

4 方法

4.1 基於統計的候選短語產生

首先將文本中特殊符號以空格替代作為斷句輔助，接著使用 jieba 初步斷詞 (Word Segmentation) 後 merge 字詞組成複合單詞，將複合單詞和 n-gram 技術產生的字詞過濾後組成候選詞庫。過濾方法包含計算 branch entropy、mutual information、word frequency。

branch entropy 以字詞左右組合字集合評斷該字詞是否容易與其他字組合成一個新詞，我們將 left entropy 和 right entropy 比較後分別設置 min entropy、max entropy 兩個閾值作為過濾條件，如 Table2 所示。

$P(w_{left}/w)$: 左鄰接字與候選字詞共同出現的機率

c_{left} : 所有左鄰接字共現率集合

$entropy(c_{left})$: 以所有左鄰接字共現率計算獲得的 left entropy

mutual information 則判斷兩個字詞相關程度決定其是否可組成一個複合詞。

$P(w_1, w_2)$: w_1, w_2 作為一組合字出現的機率

$P(w_1)$: w_1 出現的機率

$aggre_coef = \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$: 自 mutual info 延伸出的自定義係數

4.2 新短語篩選

4.2.1 基於序列標記的短語篩選

針對新短語發現相關方法所提取出的候選新短語，我們希望訓練一個基於序列標記、仿造命名實體辨識 (NER) 的神經網路模型，判斷文本、句子中的字詞特徵是否符合新短語的標準，並將該詞語標記出來。實作上我們使用哈工大開源的 BERT 模型：chinese-bert-wwm-ext(Cui et al., 2019) 對句子中的每個字生成 BERT word embedding，後續串接一層 Dropout 層和 Linear 層的神經網路作為是否為新短語、但判斷新短語類別的分類器，來預測目標短語是否為合適的新短語，另外，訓練時我們凍結了 BERT 模型的預訓練參數，只更新分類器神經網路的權重，使用的 Loss Function 為 CrossEntropyLoss。

4.2.2 基於分類與文本蘊含的短語篩選

短語篩選可以被視為一種單純的分類任務，即使在不依賴短語所屬語句脈絡的情況下，大部分短語的類別以及是否通順仍然可以被判斷。而在短語分類中加入語句脈絡有可能會藉由 BERT 的 Self-Attention 機制提升分類的效果。同時，我們還注意到，進行分類任務時如果將文本的標籤語意融合進模型中，將文本分類修改為蘊含關係判斷，可能會提升模型的表現 (Wang et al., 2021)。

因此，我們測試了一系列基於分類的方法，包括：(1) 只輸入詞語進行分類；(2) 輸入詞語和該詞所屬文章句子進行分類；(3) 輸入詞語，但是將分類任務修改為文本蘊含關係判斷；(4) 輸入句子，進行蘊含關係判斷。為了將分類任務修改為蘊含關係判斷，我們使用了如下的模板：「{} 不是一個短語」、「{} 是關於其他的短語」、「{} 是關於工人、勞動的短語」、「{} 是關於實體的短語」。{} 為候選短語。例如，在 (4) 的設定下，一個實際輸入為：

- [CLS] 建築工人正面臨老齡化問題 [SEP]
建築工人不是一個短語
- [CLS] 建築工人正面臨老齡化問題 [SEP]
建築工人是關於其他的短語
- [CLS] 建築工人正面臨老齡化問題 [SEP]
建築工人是關於工人、勞動的短語
- [CLS] 建築工人正面臨老齡化問題 [SEP]
建築工人是關於實體的短語

我們會對訓練模型對每一個輸入語句進行二元判斷，取機率最高的句子所對應的類別作為該詞的預測結果。我們使用的模型為哈工大開源的 chinese-roberta-wwm-ext(Cui et al.,

2019), 訓練過程使用 [CLS] Token 所對應的輸出 Embedding 進行線性分類。Loss Function 為 Cross Entropy。

4.3 關鍵詞抽取

我們以中文為主，實作前面有提及的非監督式與監督式方法，以準備的文本進行研究。中文的文章在進入關鍵詞擷取前，皆會使用 jieba 套件進行分詞，並在分詞前結合我們的自定義辭典幫助斷詞的正確性，並在分詞過後以空格將段落組合起來。

針對 KeyBERT 方面，我們會將 KeyBERT 選擇的前 50 筆關鍵字資料進行權重整理，如果關鍵字貼合我們的辭典名詞，則將關鍵字評分依倍率調高，進而從中選取前 10 筆高分資料，可以有效地針對領域提取關鍵字，進而，研究此種行為是否有助於在領域中的關鍵字可以更容易地被提取出，以增進效能。

5 實驗結果與討論

在這一節中我們會展示短語篩選和關鍵詞抽取部分的結果，並討論不同設定下模型表現差異的可能原因。

5.1 短語篩選

5.1.1 基於序列標記的短語篩選

在基於序列標記的短語篩選的部分，需利用 BIO 標註方法額外製作訓練及測試資料，經過統計，大部分的字詞類別皆為 O，相較於 General、Labor、Name 等短新短語類別的出現頻率，最少高出了 100 多倍以上，類別分布非常不平均。

實驗中，我們在 Loss function 上，對 Weight (class weight, 分類類別權重參數) 分為預設類別權重 (等權重) 和加權類別權重兩種方式訓練，加權類別權重使用類別 Label 出現頻率的倒數作為該類別之權重。超參數的設定上，Batch size 為 32，Learning rate 為 1e-3，Epoch 數為 20，Table 3 是我們的實驗結果。

由表格可看出，在預設權重的情況下，模型幾乎無法辨別新短語和非新短語字詞的差異，由於類別 O (非新短語字詞) 相比新短語類別字詞的出現頻率約為 100~500 倍左右，其表現不佳甚至無法訓練的最大可能原因明顯為資料標籤類別分布不均的問題。

然而，即便在加權類別權重的情況下，已試圖平衡類別分布不均問題，表現仍然差強人意。我們推測，由於所標註新短語來自前述基於統計方法提取出的候選短語，因此候選短語可能無法涵蓋，所有深度學習模型會認定為具備候選短語特徵的字詞，造成訓練上模型學習不易。

Class weight	Precision	Recall	F1-score
預設類別權重	0.00	0.00	0.00
加權類別權重	0.21	15.26	0.42

Table 3: 基於序列標記的短語篩選的實驗結果

No	Input	Task	F1
1	短語	直接分類	92.16
2	短語 + 所在句子	直接分類	91.13
3	短語 + 所在句子	文本蘊含	93.11
4	短語	文本蘊含	92.97
5	短語 + 所在句子	文本蘊含	93.50

Table 4: 基於分類與文本蘊含的短語篩選的實驗結果

就基於序列標記的短語篩選方法而言，即便使用在相關任務表現優異的 BERT 模型，效果仍然不符預期。在短語篩選上，我們認為使用基於序列標記的方法會需要較複雜的方法和技巧來訓練，並不是一個非常有效率的選擇。

5.1.2 基於分類與文本蘊含的短語篩選

Table 4 是我們的實驗結果，取三次隨機初始化後的最高分數，F1 為 Macro-F1。超參數設置基本一致，Batch Size 為 32，Learning Rate 為 1e-5，Epoch 數取 10。這張表格說明以文本蘊含方式進行模型訓練與推論能獲得較好的結果，而加入短語所在句子對於直接分類沒有幫助，但對於文本蘊含任務則有提升效果。此外，No.5 實驗中，我們在四個模板之外，額外加入了一個模板「{} 是一個短語」，標籤為 General、Labor、Name 的候選短語此處的 Ground Truth 都設置為 1。這個額外模板在推論中並沒有作用，但可以讓模型獲得額外的訊息 (即三種類別與 O 相對立)，結果提升了模型的表現。

我們認為，這樣的實驗結果證明了將簡單分類任務修改為文本蘊含作為一個訓練技巧的有效性。但需注意的是，這樣的任務修改在計算資源上消耗較多，在推論時因為要判斷的次數等於模板個數，而直接分類僅需要判斷一次，也會讓模型速度較慢。我們認為實際應用場景下，可以先設置門檻值，對「{} 是一個短語」模板進行判斷，超過閾值被認為有可能是短語的情況下再進行完整全模板的推論。在非短語較多的情況下，這樣的作法應該可以大大提升推論速度。

5.2 關鍵詞抽取

Table 5 是我們針對「上海疫情實錄：被驅逐的“橋洞”騎手，生活在地下的日結零工 | 服務業勞動觀察」這個條目去作方法比較中挑出

KeyBERT	RAKE	TextRank	YAKE
橋洞	月	騎手	地下通道
但小鑫	日	小鑫	大部分
當晚	傍晚	橋洞	志願者
朝不保夕	點	上海	越來越
騎手	晚	提供	街道辦
鐵門	橋洞	地方	救助站
徐匯區	位	住	包工頭
封城後	十	疫情	上海市
驅逐	樣	只能	礦泉水
日晚	住處	說	停車場

Table 5: 基於不同方法的關鍵字擷取實驗結果

KeyBERT	KeyBERT 加入詞典
橋洞	眾包騎手
但小鑫	橋洞
當晚	專送騎手
朝不保夕	隔離期間
騎手	朝不保夕
鐵門	但小鑫
徐匯區	小鑫聊
封城後	高架橋
驅逐	徐匯區
日晚	鐵門

Table 6: 基於 KeyBERT 與改良後的的關鍵字擷取實驗結果

來的 10 個關鍵字，直觀而言我們就可以得出成效以 KeyBERT 最佳、TextRank 及 YAKE 次之，而 RAKE 則幾乎沒有作用。

Table 6 是針對標準 KeyBERT 模型與增加我們算法改良的 KeyBERT 算法進行的成效比較。也因此從結果可以發現有關勞工部分的關鍵字得以更加靠前，進入到我們所選定的 10 個關鍵字中。

6 結論

本研究證明，預訓練語言模型結合傳統統計方法，可以在少量文本中有效地進行短語發現、篩選並製作短語詞典。對於加入詞典前後關鍵詞抽取的 Case Study 也表明，這種作法可以提高以領域詞典為基礎的下游任務的表現。但更多下游任務的量化評估仍需在公開資料集上進行更多的實驗。我們希望本研究可以提供一個結合預訓練語言模型與傳統詞袋模型方法的可能路徑。

References

- Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.
- Qiuyuan Chen, Guang Cheng, Di Li, and Jian Zhang. 2017. Closeness based new word detection method for mechanical design and manufacturing area. *電腦學刊*, 28(5):210–219.
- Anacleto Correia, M Filomena Teodoro, and Victor Lobo. 2018. Statistical methods for word association in text mining. In *Recent Studies on Risk Analysis and Statistical Modeling*, pages 375–384. Springer.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.
- Maarten Grootendorst. 2020. [Keybert: Minimal keyword extraction with bert](#).
- Kaiyu Huang, Keli Xiao, Fengran Mo, Bo Jin, Zhuang Liu, and Degen Huang. 2021. Domain-aware word segmentation for chinese language: A document-level context-aware model. *Transactions on Asian and Low-Resource Language Information Processing*, 21(2):1–16.
- Oskar Jerdhaf, Marina Santini, Peter Lundberg, Anette Karlsson, and Arne Jönsson. 2021. Implant term extraction from swedish medical records—phase 1: Lessons learned. In *Swedish Language Technology Conference and NLP4CALL*, pages 35–49.
- Qi Jia, Yonghong Xie, Cong Xu, Yue Zhou, and Dezheng Zhang. 2019. Unsupervised traditional chinese medicine text segmentation combined with domain dictionary. In *International Conference on Artificial Intelligence and Security*, pages 304–314. Springer.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.
- Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1:1–20.
- Gaohui Shang. 2019. Research on chinese new word discovery algorithm based on mutual information. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 580–584.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Jingjing Xu and Xu Sun. 2016. Dependency-based gated recursive neural network for chinese word segmentation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–572.

Jie Yang, Yue Zhang, and Shuailong Liang. 2018. Subword encoding in lattice lstm for chinese word segmentation. *arXiv preprint arXiv:1810.12594*.

Yanfang Zhang, Zhi Li, Jiaqiang Wang, Chuanhu Zhao, Yabo Xu, and Yue Ge. 2019. A new word discovery method based on constrained and dsg. In *2019 Chinese Automation Congress (CAC)*, pages 5299–5302. IEEE.