生物醫學實體檢測模型之實驗與錯誤分析
SCU-NLP at ROCLING 2022 Shared Task: Experiment and Error Analysis of Biomedical Entity Detection Model

Sung-Ting Chiou
Soochow University
Dept. of Data Science
yuki1214888@gmail.com

Sheng-Wei Huang
Soochow University
Dept. of Data Science
kiwihwung11@gmail.com

Ying-Chun Lo
Soochow University
Dept. of Data Science
ginny880530@gmail.com

Yu-Hsuan Wu
Soochow University
Dept. of Data Science
Ikaroskasane28@gmail.com

Jheng-Long Wu
Soochow University
Dept. of Data Science
jlwu@gm.scu.edu.tw

摘要

生物醫學之命名實體辨識相較於一般命名實體辨識任務來得更加複雜。本次命名實體辨識任務以辨識醫療保健領域的十種命名實體類型為目的，預測句子的命名實體邊界和類別。我們探討了命名實體辨識的多種基礎方法，如隨機森林、隱馬爾可夫模型、條件隨機場和 BERT。提供未來在醫療領域的 NER 辨識中，能選擇最佳表現的基礎方法為基準進行改良。預測結果以 BERT 模型在 F-score 上較為顯著，取得了更好的結果。

Abstract

Named entity recognition generally refers to entities with specific meanings in unstructured text, including names of people, places, organizations, dates, times, quantities, proper nouns and other words. In the medical field, it may be drug names, Organ names, test items, nutritional supplements, etc. The purpose of named entity recognition in this study is to search for the above items from unstructured input text. In this study, taking healthcare as the research purpose, and predicting named entity boundaries and categories of sentences based on ten entity types, We explore multiple fundamental NER approaches to solve this task, Include: Hidden Markov Models 、 Conditional Random Fields、Random Forest Classifier and BERT. The prediction results are more significant in the F-score of the CRF model, and have achieved better results.

關鍵字：實體命名、隱馬爾可夫模型、條件隨機場、隨機森林林

Keywords: Named entity recognition 、BERT、Random Forest Classifier 、 Hidden Markov 、 Conditional Random Field

1 緒論

命名實體辨識（Named entity recognition，NER）任務是自然語言處理的基本任務，同時也作為許多應用的基礎。例如：翻譯、文本摘要。因此進行 NER 任務能否使模型達到更好的辨識效果是許多研究所追求的。除此之外 NER 任務也根據領域的應用有所差異，且若採用監督學習也需另行標記實體的資料集。舉例而言 NER 常見的分類多為地點、時間、人名、組織等，然而應用上也能將此任務進行特定語句結構的實體辨識，例如激進言語、反諷等。本次的分享任務即是進行醫療領域的 NER 辨識。此類型的任務相當重要，NER 的效果將會影響後續任務的可靠性，因此探討如何提升醫療 NER 的辨識效果是本文的主軸。

本次的分享任務將進行醫療領域的中文 NER 辨識。中文 NER 相較於英文處理上較困難。如何正確的分辨實體的邊界也是難題之一。以此為基礎便延伸出基於字符的方法和基於單詞方法，本研究也將探討使用兩種詞

嵌入訓練的模型效果。現代在網路的資訊流通快速，用戶能夠通過網路搜尋醫療相關資訊，在進行就醫。因此網路上能產生很多醫療相關的文本，這也提供了建立醫療領域 NER 辨識的資料豐富度。自動識別醫療保健、生醫領域的實體能夠協助歸納或萃取醫療文本中的資訊。本次任務共需辨識 10 種實體類型，需預測每個給定句子的命名實體邊界和類別。使用的訓練與料庫為 Chinese HealthNER 語料庫(Lee and Lu, 2021)。包括 30,692 個句子，總計約 150 萬個字符或 91700 個單詞。經過人工註釋，有 68,460 個命名實體， 10 種實體類型分別是：身體、症狀、儀器、檢查、化學、疾病、藥物、補充劑、治療和時間。訓練資料集中包含了語句、字符、分詞的文本資料以及對應的 NER 分類(Lee et al., 2022)。

過往研究在 NER 的辨識任務上採取的策略都不同，然而在模型建構上的巧思可歸納為改良或組合多種模型，或是增加詞嵌入的資訊(如筆畫、部首等資訊)。本文旨在探討、比較基礎模型的效果以提供後續研究在改良模型時對基礎模型的選擇。本文針對機器學習方法、深度學習方法也進行了效果的比較。機器學習方法使用隨機森林，深度學習方法使用 Bidirectional Encoder representations from transformers (BERT)、隱馬爾可夫模型(Hidden Markov Model ， HMM) 、條件隨機場(conditional random field，CRF) 等三種近年 NER 辨識任務中仍常用來改良或組合的基礎模型進行比較。

2 文獻回顧

Li 等人(Li, 2020) 對以往 NER 任務的解決方法進行深入探討，介紹了傳統 NER 方法建構，以及詳細介紹了近年使用深度學習取得的成果。傳統基於規則的 NER 辨識方法依賴於字典的建構，也由於此特性特定領域的規則和不完整的字典，從此類系統中經常觀察到高精度和低召回率，並且無法將系統轉移到其他領域。非監督學習方法也有研究證明了其效果的有效性和普遍性。監督方法進行 NER 依賴特徵工程，機器學習方法。常見的方法如 HMM、決策樹等等。其中 CRF 的 NER 已廣泛應用於各個領域的文本。包括生物醫學文本 、推文和化學文本 。基於這些傳統方法，對於 NER 的研究越來越多元，也釋出許多分享任務。Lee 等人 (2020) 針對臨床命名實體識別在未標記的臨床記錄上預訓練 BERT 模型。並使用長短期記憶（LSTM）和條件隨機場（CRF）提取文本特徵和解碼預測標籤，並提出一種將字典特徵整合到模型中的新策略。Segura Bedmar 等人 (2013) 提出了兩個子任務：1.藥物名稱的識別和分類 2.相互作用的提取和分類，作為 SemEval 2013 任務 9 的一部分，其研究結果顯示，在命名實體任務中，參與系統在識別已知實體方面表現良好。Alsehaimi 等人 (2022) 使用自然語言處理 (natural language processing，NLP) 的 NER 技術在關於酒店的大量文本評論數據中找到主要實體的自動識別器。並在五種不同的分類模型如：Spacy, Naïve Bayes (NB), Stochastic Gradient Descent (SGD), Passive Aggressive,和 AdaBoost 之間進行比較，在 1000 多條記錄的真實數據集上進行實驗，結果顯示 NB 的準確率最高。

Kocaman 等人 (2021) 在 Apache Spark 之上重新實現 Bi-LSTM-CNN-Char 深度學習架構。Luo 等人 (2020) 通過標籤嵌入註意機制增強從獨立 BiLSTM 學習的句子表示。Mayhew 等人 (2020) 利用有噪聲的資料進行育訓練改善 BiLSTM-CRF 模型和 BERT 嵌入的效果。Han 等人 (2021)提出 MAF-CNER 模型，針對中文的多種特徵進行融合、訓練。Englmeier 和 Mothe (2020) 將 NER 應用於激進言語的偵測。Carbonell 等人 (2020) 使用圖神經網絡架構來解決半結構化文檔中的實體識別和關係提取問題。Wu 等人 (2021) 利用 Transformer 架構進行改良，將文字結構與字符資訊引入模型進行交叉注意力訓練。Wang 等人 (2021) 針對實體提及不連續的問題提出了解法，其模型成果優於最先進的結果， F1 上提高了 3.5 個百分點，並且實現了 5 倍的加速。Zhang 等人 (2020) 在中藥的 NER 辨識上提出了 Back-Labeling 的方法，分辨實體的跨度是否為連續，以此進行模型訓練提升效果。

Kumar 和 Starly (2021) 以 BiLSTM+CRF 的神經網絡模型建構了製造業的 NER 辨識模型，有利於製造業未來能有程序化查詢和檢索系統。Li 等人 (2021) 提出 MIN 模型，利用段級

訊息和詞級依賴關係，結合一種交互機制來支持邊界檢測和類型預測之間的信息共享。Litake 等人 (2022) 針對各種 BERT 的變體進行測試，並觀察多語言的預訓練模型與單語言的預訓練模型效果差異。此研究與本文的性質相似，本研究主旨在於將 NER 任務中常見的基礎神經模型進行效果差距的評估。Wang 等人 (2021) 認為實體檢測和關係提取模型設置兩個獨立的標籤空間可能會阻礙實體和關係之間的信息交互，因此將其融合。

綜上所述可知 NER 辨識不僅重要，且跨足多個領域都有需求。提升效果的方法更是多元，在標記階段改良、詞嵌入改良、模型組合改良皆有方法提出。近兩年的研究以神經網路進行 NER 辨識為主。因此本文旨在針對常見的神經模型 HMM、CRF 以及現今已被證實在多種 NLP 任務能提升整體任務效果的預訓練模型 BERT 之間的效果進行比較。鑒於機器學習方法進行 NER 辨識的研究仍持續精進，因此本研究也採取隨機森林(Random Forest Classifier，RFC)進行比較。

3 研究方法

本章節介紹本研究用於解決 NER 任務的四個模型，如 Random Forest Classifier 模型、HMM 模型和 CRF 模型和 BERT 模型。在將結果與分類器的預測輸出進行比較時，可能會出現不同的情況。例如，字符串匹配第二個類別，第三個預測不正確的類別。 模型檢測中缺少該類別，因此分別計算每個類別的 Precision、Recall 和 F1 分數。

3.1 Random Forest Classifier

隨機森林是一種基於決策樹的機器學習分類模型，可以根據標記的術語學習基本規則，由於具備準確性、簡單性、靈活性，使得 RFC 成為機器學習分類模型中流行的模型之一。

3.2 Hidden Markov Model

HMM 模型是一種觀察觀測值來估算狀態的模型，主要著重在觀測值的前後順序關係。也就是說，目標觀測值的前一個與後一個觀測值將會是影響估算目標狀態的主要因素。在 NLP 任務中，HMM 模型會藉由觀測目標字詞的前一個與後一個字詞，估算目標字詞之狀態。HMM 模型的優點在於能考慮字詞之間的前後關係，因為在句子當中字詞順序的確是重要的特徵之一。例如，給定一句話「她是一位女孩」，其中因為模型著重順序關係，因此「女孩」是會受到「她」影響的，這也與人們理解一般字詞之間的關係相符合。

3.3 Conditional Random Field

在前一節模型介紹的 HMM 模型主要考慮字詞順序性，然而並沒辦法呈現句子中最真實的狀態。舉例來說給定一句話「她是一位女孩」，雖然「女孩」一詞確實會受到「她」影響，但是 HMM 模型因為著重順序關係，因此模型中學習到影響「女孩」一詞最多的將會是「一位」而並非「她」，與事實有著些許不相符。而 CRF 模型相對於 HMM 模型，主要考慮的是字詞之間的相互關係，藉由計算整句句子中各字詞之間的條件機率，能學習到字詞之間最真實的關係。

3.4 BERT

BERT 是一種透過預訓練大量文本所得的文本表示模型，透過學習字與字之間的關係取得隱藏表示特徵。本研究使用 BERT 模型取得文本的隱藏表示特徵後，再經過預測分類層估算字與字之間的關係並取得最終分類結果。BERT 因為易於使用又能快速微調模型並串接各種不同任務，並且在各種自然語言處理任務中獲得良好的結果，可以說是目前 NLP 領域中最流行的模型。

4 實驗結果

本章節展示實驗資料在各個模型的實驗結果及比較，圖 1 為分別使用隨機森林與 BERT 以斷詞為輸入進行預測的結果，圖 2 為使用隨機森林、隱馬爾可夫模型(HMM)、條件隨機場(CRF)和 BERT 以字為輸入進行預測的結果，指標為 Precision、Recall 以及 F1-score。首先從表 1 中呈現了各個模型的實驗結果，本研究以分類任務中經常使用的 Precision、Recall 以及 F1-score 作為主要評估指標，並且分別列出針對 Word 及 Character 的分類結果。從表格中可以看出，不管是在 Word 還是 Character，BERT 在整體結果都能取得較好的效果，平均

Format	Model	Precision		Recall		F1-score	
		Mean	SD	Mean	SD	Mean	SD
Word	RandomForest Classifier	**0.82**	**0.14**	0.65	0.19	0.71	0.17
	BERT	0.79	0.13	**0.76**	**0.15**	**0.77**	**0.13**
Character	RandomForest Classifier	0.04	0.18	0.05	0.21	0.04	0.19
	HMM	0.59	0.18	0.65	0.14	0.61	0.15
	CRF	0.78	0.11	0.62	0.17	0.68	0.14
	BERT	**0.78**	**0.10**	**0.75**	**0.14**	**0.76**	**0.11**

表 1. 模型分數比較

F1-score 能取得 0.77 及 0.76 的成績。然而值得一提的是，以斷詞為輸入時，隨機森林的 Precision 效果較佳，推測以斷詞輸入時機器學習模型能較好的對照出正確答案，然而整體效果仍無法超越深度學習模型。這樣的結果也顯示，NER 任務使用機器學習的方式來解還是稍顯不足，透過深度學習的方式來學習字與字之間的隱藏關係能更有效的提升分類效果，讓模型達到精準分類的效果，此外預訓練也有助於效果得提升。

接續上述從圖 1 可以明顯看出若以斷詞為輸入進行命名實體預測時，深度學習模型相較於機器學習的效果在 F1-score 差距是明顯的。然而可以發現共通性是在醫療器具(INST)、藥物(DRUG)和治療(TREAT)的命名實體辨識上表現較差。這三種實體有較多專有名詞，可見需要特別處理，例如建立專有名詞的字典等等。

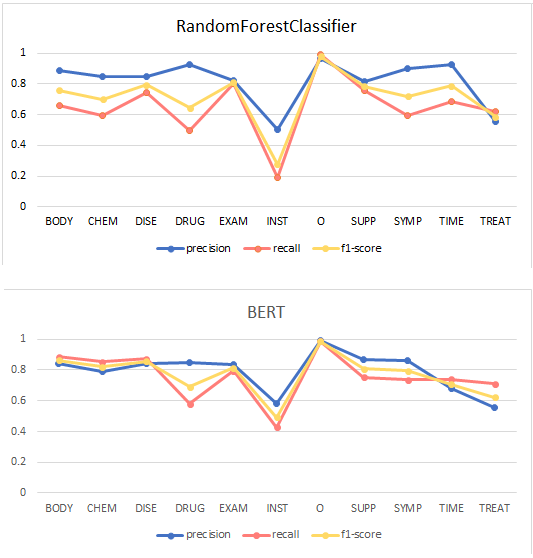

圖 1. 以斷詞為輸入隨機森林與 BERT 的結果

圖 2 可以看出隨機森林以斷字為輸入時無法進行學習，因此將所有分類都分為 O。隱馬爾可夫模型(HMM)、條件隨機場(CRF)和 BERT 三種模型在各實體的分類成效上有相似的趨勢。整體而言在所有實體分類的 F1-score 上 BERT 都優於 HMM 和 CRF。可知綜合性能上 BERT 相當出色。然而 Precision 則是 CRF 的優勢，CRF 模型的精準度得到與 BERT 相似的成果。值得一提的是 HMM 模型在治療(TREAT)和時間(TIME)的 Recall 取得與 BERT 相似的效果，推測為 HMM 訓練時考慮前後字的特性，因此在這兩種跨度較長的實體上表現較佳。

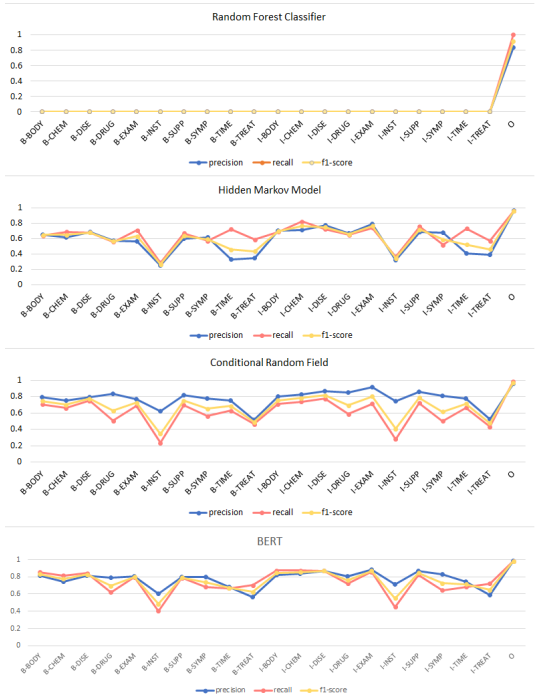

圖 2. 使用隨機森林、隱馬爾可夫模型、條件隨機場和 BERT 對斷字進行預測的 f1-score 結果

4.1 錯誤分析

針對 BERT 預測結果進行錯誤分析，本研究發覺，以詞為輸入比以字為輸入的預測效果佳。因此以下分析為，「以詞為輸入」時，特定詞的實體預測正確，然而「以字為輸入」時對應特定詞的個別字符實體預測錯誤的分析。舉例而言特定詞「尿糖」的正確分類為病徵(SYMP)，以詞為輸入的 BERT 模型預測也為病徵(SYMP)。然而以字為輸入的 BERT 模型預測會將「尿」預測為 B-BODY，「糖」預測為 I-BODY。簡言之以字為輸入的 BERT 模型將「尿糖」分類為身體(BODY)實體。類似的錯誤案例有共有 642 個詞實體預測正確，這些詞各自對應的字共 1030 個皆實體預測錯誤。推測此現象原因為，單一中文字提供的資訊較多的案例為身體器官，且實際案例中腎、肝、脾、胃…等等單一中文字就能構成器官，然而病徵、疾病、藥物…等其餘實體皆須整個詞才能構成完整的意義。

在字的實體預測中，將實體為 O 的字誤判的案例共有 314 個。其中誤判為 B 的案例共有 172 個，誤判為 I 的案例有 142 個。由於字級別能夠組合的資訊多樣而導致錯誤，這也是中文字的特性。舉例來說「細針」的實際實體為器材(INST)，然而模型的預測可能受到上下文影響將「細針穿刺」判斷為檢察(EXAM)，可見以字為輸入的模型在判斷組合上較為精細、彈性，表現出對上下文的適應，但實體的邊界可能較難掌握。類似的案例中，原文「也可以按摩血海穴來消除浮腫的身體」，其中「按摩」正確實體為 O 然而模型判斷「按」為 B-TREAT，「摩」為 I-TREAT。模型對於按摩多將其分類為治療(TREAT)，可見模型較難掌握特定名詞在何時屬於治療實體何時不屬於治療實體。

儘管以詞為輸入的模型表現較佳，本研究發現以字為輸入的模型效果差距並不大，推測以字為輸入的缺點在於邊界的資訊較難訓練。本研究在表 2 上呈現出各類別的跨度分類錯誤。篩選條件為，當以詞為輸入的模型預測正確的情況下，以字為輸入的模型預測的位置錯誤，在各類別上錯的個數。可以發現在 BODY 實體的錯誤較多，其中 I 的判斷錯誤較 B 多，原因可延續上述類別分類錯誤的問題，人的身體器官可能會混在病徵或是檢查內，因此可能在專有名詞中被錯誤的分類，或是造成位置的判斷錯誤。整體而言在 I 的位置錯誤上較多可以看出模型在跨度的判斷上表現較差，這也應證了目前 NER 任務最佳模型的方法都會混合字、詞的嵌入進行訓練的原因。

類別	B 位置錯誤	I 位置錯誤
BODY	57	78
SYMP	14	46
DISE	9	10
EXAM	3	4
CHEM	11	31
TREAT	2	0
TIME	0	0
INST	0	0
SUPP	2	0
DRUG	0	2

表 2.以字為輸入的模型跨度錯誤統計

5 結論與未來目標

總結本研究工作，本研究針對這次分享任務進行了目前 NER 任務中常用來改良或組合的基礎模型進行了表現的分析。其中以 BERT 的表現最佳，此結果的原因推測與 BERT 做的預訓練有關。BERT 預訓練中進行的遮罩訓練推測能提升模型針對跨度預測的表現，因此未來研究中能使用 BERT 作為基礎應用在字、詞嵌入或組合模型能有效提升效果。在字、詞嵌入的比較上，本研究發覺以字為輸入的模型在跨度上表現較以詞為輸入的模型稍差，差距體現在實體跨度的判斷上，然而以字為輸入的模型表現出了考慮上下文資訊的特性，能夠針對字詞的組合有彈性的判斷，因此本研究認為在判斷實體的研究中若要組合上下文的資訊進行判斷時採用以字為輸入應能提升模型效果。延續上述，本研究認為未來在 NER 任務的解法上，使用 BERT 為基礎改良，並且結合詞、字兩種嵌入作為輸入並考慮上下文的方法應能使整體效果更加提升。

References

ALSEHAIMI, Afnan Abdulrahman A., et al. A Smart Framework to Analyze Hotel Services after COVID-19. In: 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2022. p. 415-419.

Baum, L. E.; Petrie, T. (1966). "Statistical Inference for Probabilistic Functions of Finite State Markov

Chains". The Annals of Mathematical Statistics. 37 (6): 1554–1563. doi:10.1214/aoms/1177699147

Carbonell, M., Riba, P., Villegas, M., Fornés, A., & Lladós, J. (2021, January). Named entity recognition and relation extraction with graph neural networks in semi structured documents. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 9622-9627). IEEE.

Englmeier, K., & Mothe, J. (2020, July). Application-oriented approach for detecting cyberaggression in social media. In *International Conference on Applied Human Factors and Ergonomics* (pp. 129-136). Springer, Cham.

Han, X., Zhou, F., Hao, Z., Liu, Q., Li, Y., & Qin, Q. (2021). MAF-CNER: A Chinese named entity recognition model based on multifeature adaptive fusion. *Complexity*, *2021*.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805. 2018

Kocaman, V., & Talby, D. (2021, January). Biomedical named entity recognition at scale. In *International Conference on Pattern Recognition* (pp. 635-646). Springer, Cham.

Kumar, A., & Starly, B. (2021). "FabNER": information extraction from manufacturing process science domain literature using named entity recognition. *Journal of Intelligent Manufacturing*, 1-15.

L. Breiman, "Random forests. Machine learning," 45(1),pp. 5-32. (2001)

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Li, F., Wang, Z., Hui, S. C., Liao, L., Song, D., Xu, J., ... & Jia, M. (2021, August). Modularized interaction network for named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 200-209).

Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50-70.

Li, X., Zhang, H., & Zhou, X. H. (2020). Chinese clinical named entity recognition with variant neural structures based on BERT methods. *Journal of biomedical informatics*, *107*, 103422.

Litake, O., Sabane, M., Patil, P., Ranade, A., & Joshi, R. (2022). Mono vs multilingual BERT: A case study in hindi and marathi named entity recognition. *arXiv preprint arXiv:2203.12907*.

Lung-Hao Lee, and Yi Lu (2021). Multiple Embeddings Enhanced Multi-Graph Neural Networks for Chinese Healthcare Named Entity Recognition. IEEE Journal of Biomedical and Health Informatics, 25(7): 2801- 2810.

Lung-Hao Lee, Chao-Yi Chen, Liang-Chih Yu, and Yuen-Hsien Tseng. 2022. Overview of the ROCLING 2022 shared task for Chinese healthcare named entity recognition. In Proceedings of the 34th Conference on Computational Linguistics and Speech Processing.

Luo, Y., Xiao, F., & Zhao, H. (2020, April). Hierarchical contextualized representation for named entity recognition. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 05, pp. 8441-8448).

Mayhew, S., Nitish, G., & Roth, D. (2020, April). Robust named entity recognition with truecasing pretraining. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 05, pp. 8480-8487).

SEGURA-BEDMAR, Isabel; MARTÍNEZ FERNÁNDEZ, Paloma; HERRERO ZAZO, María. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013).

Wang, Y., Sun, C., Wu, Y., Zhou, H., Li, L., & Yan, J. (2021). UniRE: A unified label space for entity relation extraction. *arXiv preprint arXiv:2107.04292*.

Wang, Y., Yu, B., Zhu, H., Liu, T., Yu, N., & Sun, L. (2021). Discontinuous named entity recognition as maximal clique discovery. *arXiv preprint arXiv:2106.00218*.

Wu, S., Song, X., & Feng, Z. (2021). Mect: Multi-metadata embedding based cross-transformer for chinese named entity recognition. *arXiv preprint arXiv:2107.05418*.

Zhang, D., Xia, C., Xu, C., Jia, Q., Yang, S., Luo, X., & Xie, Y. (2020). Improving distantly-supervised named entity recognition for traditional Chinese medicine text via a novel back-labeling approach. *IEEE Access*, *8*, 145413-145421.