The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 241-245


# MONPA: 中文命名實體及斷詞與詞性同步標註系統

葉文照　Wen-Chao Yeh
臺北醫學大學大數據科技及管理研究所
Graduate Institute of Data Science
Taipei Medical University
m946107004@tmu.edu.tw

謝育倫　Yu-Lun Hsieh
中央研究院及國立政治大學
SNHCC, TIGP, Academia Sinica & National Cheng Chi University
morphe@iis.sinica.edu.tw

張詠淳　Yung-Chun Chang
臺北醫學大學大數據科技及管理研究所
Graduate Institute of Data Science
Taipei Medical University
changyc@tmu.edu.tw

許聞廉　Wen-Lian Hsu
中央研究院資訊科學研究所
Institute of Information Science
Academia Sinica
hsu@iis.sinica.edu.tw

## 摘要

有鑑於現今國內外研究繁體中文自然語言處理缺乏合適的斷詞、詞性標註及命名實體辨識的工具，本研究基於 BERT 模型，搭配 CRF 提出以多目標命名實體辨識與詞性標註(Multi-Objective NER POS Annotator, MONPA)系統，並以供學術使用授權條款 CC BY-NC-SA 4.0 License 進行相關安裝套件釋出作業。透過 MONPA 的釋出，嘉惠我國相關學術研究，俾能加快繁體中文自然語言處理之進展。

## Abstract

In view of the lack of suitable word segmentation, part-of-speech tagging and named entity recognition tools in the traditional Chinese natural language processing. This study is based on the BERT model with CRF to propose a multi-objective named-entity and part-of-speech annotator, which called MONPA. Our work not only propose a method but also release the relevant python package with the CC BY-NC-SA 4.0 License. We firmly believe that this research project can bridge the technical gap between academia and business applications with our innovation, and enable efficient development of traditional Chinese NLP by all entities in order to enhance our level of competitiveness in the world.



關鍵詞：中文斷詞, 詞性標註, 命名實體辨識, BERT

Keywords: Chinese Word Segmentation, POS tagging, Name Entity Recognition, BERT


## 一、緒論

綜觀目前繁體中文的斷詞工具主要仰賴 Jieba[1]套件，然而 Jieba 是基於簡體中文語料透過 HMM [1]模型所訓練出來的成果，因此對繁體中文的支援效果不佳，且系統多年未更新。種種的限制讓國內學界或是產業界想要進行繁體中文自然語言處理之研究困難重重。此外，命名實體辨識(named entity recognition)有助於瞭解句子結構進而提升理解能力，但在目前處理繁體中文時尚無可用的工具。繁體中文自然語言處理的基礎設施於此種種的限制之下，勢必使得臺灣的研發能力在這波 AI 浪潮中受阻。有鑑於此，本研究以深度學習方法研發一種能同時完成「命名實體辨識」、「繁體中文斷詞」以及「詞性標註」之系統，並將其完全開源釋出，讓所有想要處理繁體中文的產學界使用者共享此研究成果。

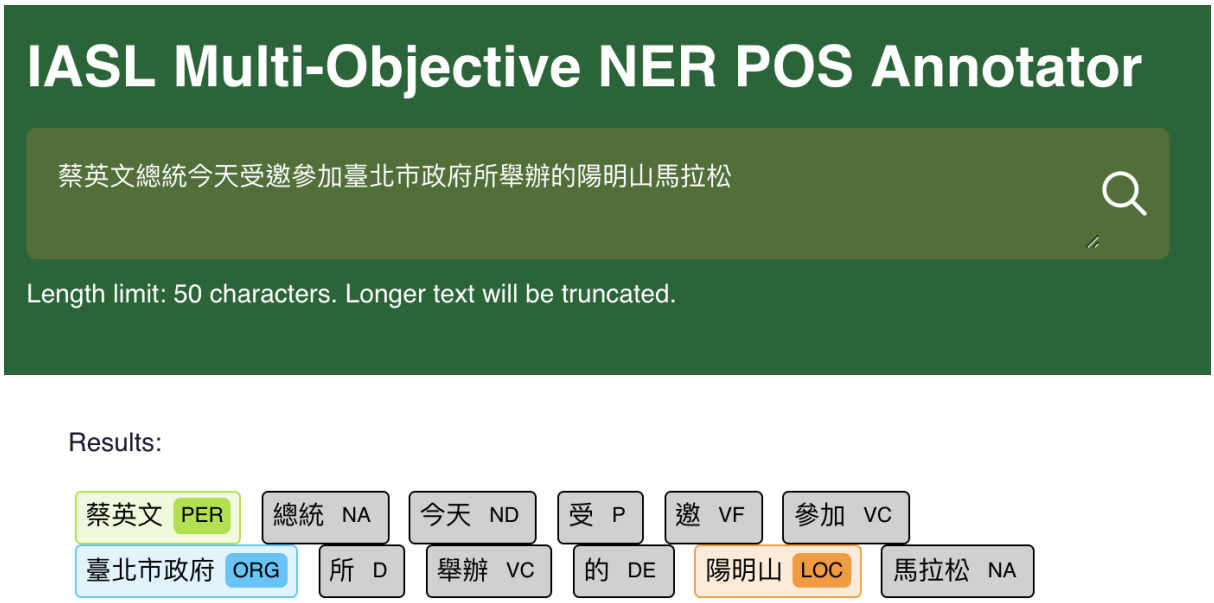
本研究所提出的多目標命名實體辨識與詞性標註(Multi-Objective NER POS Annotator, MONPA)系統，是基於 BERT [2]（應用雙向 Transformer）模型來取得更強健的詞向量（word embeddings）並配合 CRF 同時進行斷詞、詞性標註、及 NER 等多個目標。BERT 模型為現今頂尖的詞向量獲取方法之一，其利用自注意力（self-attention）機制及預訓練（pre-training）等技術以提取更能充分代表整個語句訊息的向量。本研究以授權條款 CC BY-NC-SA 4.0 License 進行相關套件釋出作業。為了使用的便利性，我們

[1]https://github.com/fxsjy/jieba

也將把 MONPA 組成套件發佈到 PyPI，讓使用者能夠透過 pip install 指令安裝。使用者可以透過 Github 獲得 MONPA 相關資訊，進而完成安裝，允許獲得 MONPA 的人依照同一授權條款的情形下再散布。透過 MONPA 的釋出，嘉惠我國相關研究與產業，俾能加快繁體中文自然語言處理之進展。

## 二、使用 MONPA

MONPA [3]是一個提供繁體中文分詞、詞性標註以及命名實體辨識的多任務模型，初期只有使用原始模型（v0.1）的網站版本[2]（如圖一）。透過本研究的釋出，MONPA 已經包裝成可以 pip install 的 python 套件包，在本次釋出中，我們也透過 BERT 改善 MONPA 的效能(v0.2)，並且發佈在 Github[3]與 PyPI[4]上。使用者能夠在不同的作業平台上透過 pip install 指令完成安裝程序，此外，本研究為了因應 pip 對套件檔案大小的限制，所以在首次引入套件時才會啟動下載最新的 model 檔。




圖一、MONPA v0.1 網頁版示範圖

本研究的釋出包含了三個功能：

- ***斷詞(cut function)***：若只需要中文分詞結果，請使用 cut 功能，回傳值是 list 格式。

[2]http://monpa.iis.sinica.edu.tw:9000/chunk

[3]https://github.com/monpa-team/monpa

[4]https://pypi.org/project/monpa/

程式及輸出如下：

```
1. monpa.cut("蔡英文總統今天受邀參加台北市政府所舉辦的陽明山馬拉松比賽。")
2. ['蔡英文', '總統', '今天', '受', '邀', '參加', '台北市政府', '所', '舉辦', '的', '陽明山', '馬拉松', '比賽', '。']
```

- ***詞性標註(pseg function)***：若需要中文分詞及該詞的 POS 標註，請使用 pseg 功能，回傳值是 list of list 格式，程式及輸出如下：

```
1. monpa.pseg("蔡英文總統今天受邀參加台北市政府所舉辦的陽明山馬拉松比賽。")
2. [['蔡英文', 'PER'], ['總統', 'Na'], ['今天', 'Nd'], ['受', 'P'], ['邀', 'VF'], ['參加', 'VC'], ['台北市政府', 'ORG'], ['所', 'D'], ['舉辦', 'VC'], ['的', 'DE'], ['陽明山', 'LOC'], ['馬拉松', 'Na'], ['比賽', 'Na'], ['。', 'PERIODCATEGORY']]
```

- ***加詞(load_userdict function)***：在 MONPA 元件中，我們提供使用者自訂詞彙的功能，透過 load_userdict function 可以將使用者詞彙檔匯入，請依『詞語 詞頻 詞性』順序製作自訂詞典文字檔。

```
1. 受邀 100 V
```

當要使用自訂詞時，請於執行分詞前先 load_userdict，將自訂詞典載入到 monpa 模組。使用 pseg function 測試，可發現回傳值已依自訂詞典分詞，譬如『受邀』為一個詞而非先前的兩字分列輸出。

```
1. monpa.load_userdict("./userdict.txt")
2. monpa.pseg("蔡英文總統今天受邀參加台北市政府所舉辦的陽明山馬拉松比賽。")
3. [['蔡英文', 'PER'], ['總統', 'Na'], ['今天', 'Nd'], ['受邀', 'V'], ['參加', 'VC'], ['台北市政府', 'ORG'], ['所', 'D'], ['舉辦', 'VC'], ['的', 'DE'], ['陽明山', 'LOC'], ['馬拉松', 'Na'], ['比賽', 'Na'], ['。', 'PERIODCATEGORY']]
```

## 三、結論

MONPA 提供繁體中文自然語言處理一個全新的分詞、詞性標註暨命名實體辨識模型，從原始的網頁版進化到現今以 Open Source 釋出的套件版，可以看到全然不同的使用效率及應用效益。套件版於釋出前已經近千萬條短文句的處理測試，並於台灣 NLP 研究圈公開後，四天內已逾 6 百多次的安裝數，Github 專案也收到超過 40 多顆星星的鼓勵。相信本研究及釋出的安裝套件必定能嘉惠我國相關研究與產業，加快繁體中文自然語言處理之進展。

## 致謝

在此感謝中央研究院中文詞知識庫小組的協助。MONPA 在經中央研究院中文詞知識庫小組同意下，使用 CKIP 斷詞元件[4]輔助製作初期訓練資料。

## 參考文獻

[1] Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The annals of mathematical statistics, 41(1), 164-171.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[3] Hsieh, Y. L., Chang, Y. C., Huang, Y. J., Yeh, S. H., Chen, C. H., & Hsu, W. L. (2017, November). MONPA: Multi-objective Named-entity and Part-of-speech Annotator for Chinese using Recurrent Neural Network. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 80-85).

[4] Ma, Wei-Yun and Keh-Jiann Chen, 2003, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff", Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pp168-171.。