門控圖序列神經網路之中文健康照護命名實體辨識

Gated Graph Sequence Neural Networks for Chinese Healthcare Named Entity Recognition

盧毅 Yi Lu, 李龍豪 Lung-Hao Lee
國立中央大學電機工程學系
Department of Electrical Engineering
National Central University
ericst91159@gmail.com、lhlee@ee.ncu.edu.tw

摘要

命名實體辨識任務的目標是從非結構化的輸入文本中，抽取出關注的命名實體，例如：人名、地名、組織名、日期、時間等專有名詞，擷取的命名實體，可以做為關係擷取、事件偵測與追蹤、知識圖譜建置、問答系統等應用的基礎。機器學習的方法將其視為序列標註問題，透過大規模語料學習標註模型，對句子的各個字元位置進行標註。我們提出一個門控圖序列神經網路 (Gated Graph Sequence Neural Network, GGSNN) 模型，用於中文健康照護領域命名實體辨識，我們整合詞嵌入以及部首嵌入的資訊，建構多重嵌入的字嵌入向量，藉由調適門控圖序列神經網路，融入已知字典中的命名實體資訊，然後銜接雙向長短期記憶類神經網路與條件隨機場域，對中文句子中的字元序列標註。我們透過網路爬蟲蒐集健康照護相關內容的網路文章以及醫療問答紀錄，然後隨機抽取中文句子做人工斷詞與命名實體標記，句子總數為 30,692 句 (約 150 萬字/91.7 萬詞)，共有 68,460 命名實體，包含 10 個命名實體種類：人體、症狀、醫療器材、檢驗、化學物質、疾病、藥品、營養品、治療與時間。藉由實驗結果與錯誤分析得知，我們提出的模型達到最好的 F1-score 75.69%，比相關研究模型 (BiLSTM-CRF, BERT, Lattice, Gazetteers 以及 ME-CNER)表現好，且為效能與效率兼具的中文健康照護命名實體辨識方法。

關鍵詞：命名實體辨識、圖神經網路、資訊擷取、健康資訊學

Abstract

Named Entity Recognition (NER) focuses on locating the mentions of name entities and classifying their types, usually referring to proper nouns such as persons, places, organizations, dates, and times. The NER results can be used as the basis for relationship extraction, event detection and tracking, knowledge graph building, and question answering system. NER studies usually regard this research topic as a sequence labeling problem and learns the labeling model through the large-scale corpus. We propose a GGSNN (Gated Graph Sequence Neural Networks) model for Chinese healthcare NER. We derive a character representation based on multiple embeddings in different granularities from the radical, character to word levels. An adapted gated graph sequence neural network is involved to incorporate named entity information in the dictionaries. A standard BiLSTM-CRF is then used to identify named entities and classify their types in the healthcare domain. We firstly crawled articles from websites that provide healthcare information, online health-related news and medical question/answer forums. We then randomly selected partial sentences to retain content diversity. It includes 30,692 sentences with a total of around 1.5 million characters or 91.7 thousand words. After manual annotation, we have 68,460 named entities across 10 entity types: body, symptom, instrument, examination, chemical, disease, drug, supplement, treatment, and time. Based on further experiments and error analysis, our proposed method achieved the best F1-score of 75.69% that outperforms previous models including the BiLSTM-CRF, BERT, Lattice, Gazetteers, and ME-CNER. In summary, our GGSNN model is an effective and efficient solution for the Chinese healthcare NER task.

Keywords: Named Entity Recognition, Graph Neural Networks, Information Extraction, Health Informatics

致謝 Acknowledgements

This work was partially supported by the Ministry of Science and Technology, Taiwan under the grant MOST 108-2218-E-008-017-MY3.