The 2018 Conference on Computational Linguistics and Speech Processing
ROCLING 2018, pp. 140-141

未登錄詞之向量表示法模型於中文機器閱讀理解之應用

An OOV Word Embedding Framework for Chinese Machine Reading Comprehension

羅上堡 Shang-Bao Luo
國立臺灣科技大學資訊工程系
Department of computer science and information engineering
National Taiwan University of Science and Technology
M10615012@mail.ntust.edu.tw

李青憲 Ching-Hsien Lee
工業技術研究院巨資中心
Computational Intelligence Technology Center
Industrial Technology Research Institute
C.H.Lee@itri.org.tw

陳冠宇 Kuan-Yu Chen
國立臺灣科技大學資訊工程系
Department of computer science and information engineering
National Taiwan University of Science and Technology
kychen@mail.ntust.edu.tw

摘要

在使用深度學習(Deep Learning)方法於自然語言處理的問題時，通常會先將每一個詞以一個相對應的詞向量(Word Embedding)表示，再輸入至各式神經網路模型。當遭遇未登錄詞(Out-of-Vocabulary, OOV)的問題時，常見的處理方式是略去該未登錄詞、以一個零向量表示或是用一個隨機產生的向量表示這個未登錄詞。就我們所知，在目前的研究裡，似乎仍未有一套合理且快速的做法，用於產生未登錄詞的詞向量表示法，並進一步地探討未登錄詞的詞向量對於各式任務成效的影響性。因此，本論文嘗試提出一套新穎的詞向量表示法學習技術，其目標是為未登錄詞產生一個較為合理且可靠的低維度向量表示法；除此之外，本研究進一步地把此一技術運用於中文機器閱讀理解任務之中，探究未登錄詞對於中文機器閱讀理解任務之影響，並驗證本論文所提出的詞向量表示法學習技術之成效。

關鍵詞：深度學習、詞向量表示法、未登錄詞、機器閱讀理解。

致謝

This research was partially supported by the Project H367B83300 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan.