学习挑选伪标签：一种用于命名实体识别的半监督学习方法

李真真，冯大为，李东升，卢锡城
国防科技大学计算机学院，中国长沙市，410073

摘要：深度学习模型在命名实体识别（NER）中实现了最先进的性能；然而，其良好性能很大程度上依赖于大量标记数据。在某些特定领域，例如医学、金融和军事领域，标记数据非常稀缺，而未标记数据则很容易获得。过往研究使用未标记数据丰富词的表示，却忽略了未标记数据中对NER任务很可能有帮助的大量实体信息。本文提出一种用于NER任务的半监督方法，其通过学习一个判别模块筛除错误伪标签，以创建高质量标注数据。伪标签是为未标记数据自动生成的标签，并被当作真实标签用来训练模型。该半监督框架包括3个步骤：为特定NER任务构建最佳单神经网络模型，学习一个评价伪标签的模块，以及迭代创建新的标记数据和改进NER模型。两个英语NER任务和一个中文医疗命名实体识别任务的实验结果表明，该方法进一步提高了最佳单神经模型的性能。当仅使用预训练的静态词嵌入且不依赖任何外部知识时，该方法可获得与CoNLL-2003和OntoNotes 5.0英语NER任务上最先进模型相当的性能。

关键词：命名实体识别；无标注数据；深度学习；半监督学习方法
https://doi.org/10.1631/FITEE.1800743