標題: 深度學習在藥物警戒方面的應用：用於標記推特帖子中不良藥物反應的循環神經網路架構

摘要

目標：社交媒體是重要的藥物警戒數據來源，用於識別不良藥物反應（ADR）。由於數據量巨大，人工審查社交媒體數據變得不可行，因此需要自然語言處理技術。社交媒體包括非正式詞彙和不規則語法，這挑戰了自然語言處理方法。我們的目標是開發一種可擴展的深度學習方法，以超越社交媒體中ADR檢測性能的最新技術水平。

材料與方法：我們開發了一個循環神經網路（RNN）模型，用於標記輸入序列中的單詞，並附上ADR成員標籤。唯一的輸入特徵是單詞嵌入向量，可以通過與任務無關的預訓練或ADR檢測訓練過程中形成。

結果：我們表現最佳的循環神經網路（RNN）模型使用了從大型非特定領域的Twitter數據集中創建的預先訓練的單詞嵌入。與基準詞彙系統的F-measure為0.631和最先進的條件隨機場模型的0.65相比，該模型在數據集上實現了ADR識別的近似匹配F-measure為0.755。特徵分析表明，預訓練的單詞嵌入中的語義信息增強了靈敏度，並與循環神經網路中捕獲的上下文感知力相結合，提高了精確度。

討論：我們的模型無需特定任務的特徵工程，這表明可以應用於其他附加的序列標記任務。學習曲線分析顯示，相比其他模型，我們的模型在較少的訓練示例下達到了最佳性能。

結論：使用具有上下文感知模型和來自大型未標記數據集的詞嵌入，可以顯著提高社交媒體中的ADR檢測性能。這種方法減少了手動數據標記的要求，並且可以擴展應用於大型社交媒體數據集。

背景和意義

對於不良藥物反應（ADR）的全面了解可以減少其對患者和醫療系統的不良影響。實際上，臨床試驗無法調查藥物將被使用的所有環境，這使得在藥物獲批准之前完全描述其不良反應效應變得不可能。藥物警戒（或稱市場後藥物安全監測）是指藥物上市後對其進行的不良反應監測。大多數目前的藥物警戒活動依賴於被動的自發報告系統數據庫，如美國FDA（聯邦藥物管理局）的不良事件報告系統（FAERS）。這些系統可能會受到事件報告的延遲、偏見和不足報告的限制。例如，向FAERS報告嚴重不良事件的報告率估計僅為實際事件的1-13％。

為了解決被動報告系統的局限性，主動的藥物警戒方法不斷地分析經常更新的數據來源。最初的研究研究了從結構化電子健康記錄（EHR）數據（例如問題清單）中提取ADR的方法。然而，這些研究顯示，結構化的EHR數據包含有限的ADR信息，因此隨後的研究檢查了EHR臨床敘述中的ADR。更近期的研究將社交媒體文本作為EHR數據的補充進行研究。Twitter特別有趣，因為它具有龐大的用戶群、人口統計變異性和公開可用的數據。此外，已經確定了在Twitter數據中描述的某些ADR與在FAERS中報告的ADR之間存在顯著的相關性（P < 0.001），這表明Twitter是可行的藥物警戒數據來源。

在社交媒體中進行ADR檢測需要自動化方法來處理大量的數據。最早的研究使用字符串匹配方法來識別與預定義的藥物和不良事件詞彙相匹配的術語。然而，詞彙匹配無法區分與藥物相關的事件是對藥物的反應還是對藥物的指示。社交媒體語言的其他特徵進一步限制了詞彙匹配作為ADR檢測方法。

例如，社交媒體語言是非正式的，包括來自白話的詞彙和短語（例如“感覺像垃圾”），拼寫錯誤頻繁，文法不規則（例如“這藥真讓我出汗”）。它還包括縮寫（例如“lol”）和符號（例如表情符號），以傳達語義信息。為了應對這些挑戰，一些研究人員將ADR檢測視為監督式機器學習的序列標記問題，學習方法可以考慮給定輸入詞的周圍上下文。這在自然語言處理中很常見，其中每個標記（即連續的字符序列，通常對應一個詞）都被標記為一個命名實體標籤（例如人名）。對於ADR檢測，標記可以作為不良事件的一部分。最成功的ADR序列標記方法使用條件隨機場（CRF）模型。CRF受到其輸入範圍的限制，該模型僅考慮目標單詞及其在固定寬度窗口內的鄰近單詞，因此可能排除遠處提及的重要信息

递归神经网络（RNNs）可以克服这个限制。RNNs将文本作为一系列单词进行处理，并包含一个“记忆”任意数量先前标记决定的隐藏状态，用于标记当前标记。这种“记忆”使得RNNs非常适合标记任务，其中个别单词的标签取决于单词本身以及在其之前的单词和标签。RNN模型的变体在词性标注和命名实体识别方面取得了最先进的性能。RNNs已用于从纵向临床数据中阐明患者表型，但据我们所知，尚未特定应用于标记药物事件，尤其是社交媒体数据。

目標

我們試圖開發一個超越目前社交媒體中ADR檢測最先進技術的RNN模型。我們研究了一種特定架構，稱為雙向長短期記憶（BLSTM）RNN。在標準RNN模型中，網絡的記憶均依賴於所有先前的輸出。在LSTM變體中，模型為每個序列項目學習其先前輸出和當前輸入的權重。此外，BLSTM以前向和後向方向處理序列，使其能夠學習雙向的依賴關係。我們將幾種BLSTM變體的性能與最先進的CRF模型和基於詞彙的基準方法進行了比較。我們具體考慮了Twitter用戶發文的ADR標記性能。

材料和方法

這項研究被確定為費城兒童醫院的機構評審委員會（Institutional Review Board）審查豁免。

資料收集和標記

我們的資料集合併了兩個 Twitter 資料集。第一個資料集是 Twitter ADR 資料集（v1.0），此前已發表。21 資料集的作者使用廣泛在美國市場上使用的 81 種藥物和 2007 年至 2010 年間新推出的藥物的搜索詞來收集和註釋用戶發表的帖子（稱為推文）。20,21 公開的資料包括獨特的推文識別符，但不包括推文文本（根據 Twitter 應用程式接口許可協議的規定禁止包括在內）。因此，需要使用這些識別符從 Twitter 應用程式接口獲取文本。在原始資料集中有 957 個識別符，本研究進行時，可以下載 641 條推文。這些推文所代表的藥物並非針對特定疾病條件收集，而是代表潛在不良反應的廣泛範疇。這些資料佔我們完整資料集的 76%。

我们增补了这个数据集，加入了额外的推文，称为 ADHD 数据集，其中至少包含 44 种用于治疗注意力缺陷多动障碍（ADHD）的品牌和通用药物的 1 个搜索词。我们在 2015 年 5 月 1 日至 2015 年 12 月 31 日期间收集了这些推文。排除了带有 URL 和被转发的推文。从中，我们包括了 203 条被确定为包含与药物相关事件的推文，并使用与 Twitter ADR 数据集（v1.0）一起发布的指南进行了标注。标记过的 ADHD 数据集可以在 <https://github.com/chop-dbhi/twitter-adr-blstm> 找到。

我们的完整数据集包含 844 条推文，其中 95% 的推文至少包含一个 ADR 或指示提及。我们保持了与 Twitter ADR 数据集一起发布的训练-测试拆分，随机分割了 ADHD 数据集以创建大小为 634（75%）和 210（25%）推文的训练和测试集。训练推文中包含 647 个 ADR 和 71 个指示，而测试推文中包含 199 个 ADR 和 22 个指示。

我们使用 ark-twokenize-py Python 模块将推文分割为单独的标记。我们将所有标记转换为小写，并用特殊标记替换“at”提及（即 @用户名）。尽管 RNN 模型可以处理任意长度的序列，但许多建模框架期望批处理处理的输入具有固定长度。因此，为了实现方便，有必要标准化每条推文中的标记数。因此，我们对每条推文进行了填充，以使其长度与训练数据中最长推文的长度相匹配（36 个标记）。对于具有 k 个标记的推文，如果 k < 36，则我们在推文开头添加了 (36 - k) 个“<PAD>”标记，其中每个 <PAD> 标记是零向量。如果测试数据中的推文标记超过 36 个，则设置模型将其截断为仅包括前 36 个标记。事实证明，没有测试推文包含超过 36 个标记。

必须手动标记 ADHD 数据集中的推文，标记其是否包含药物事件为正面或负面。对于事件为正面的情况，标记了事件类型和包含事件的文本范围。两位作者独立标记了所有推文。我们对不一致的评估进行了共识调解，直到达成100%一致意见。发布的 Twitter ADR 数据集包含相同的标签。为了便于监督模型的训练和评估，需要生成指示 ADR 成员资格的标记。序列标记任务的标准方法是使用 I-O-B 方案，其中分配标签以指示标记在感兴趣实体的开始（B）、内部（I）或外部（O）的位置。我们采用了一个带有 4 个标签的 I-O 方案：IADR、I-Indication、O 和 <PAD>，分别表示给定标记是 ADR 的一部分、指示的一部分、ADR 或指示的外部，或者是填充（图 1）。这些标记可以根据事件类型和文本范围标签自动生成。

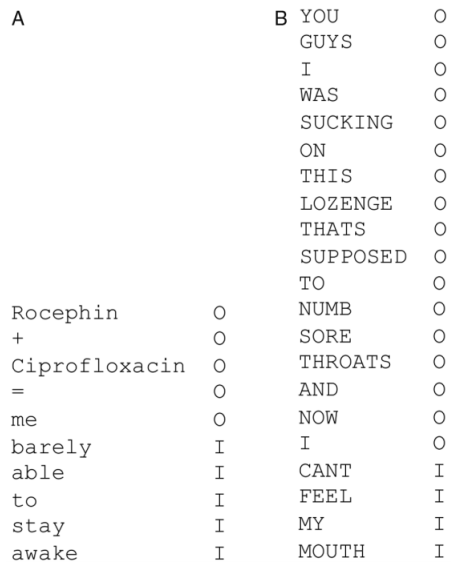


圖1. 使用我們的二元標籤集標記ADR跨距的推文範例

BLSTM-RNN 模型的开发

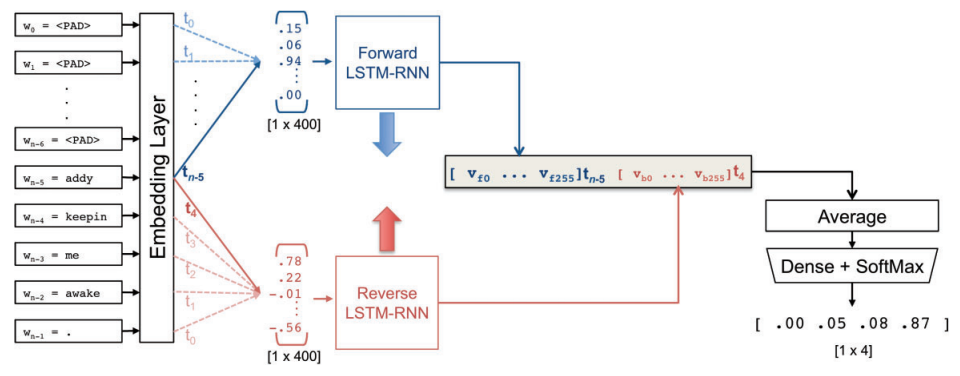
我們的雙向長短期記憶遞歸神經網路（BLSTM-RNN）模型將每個推文視為一系列的標記，並預測每個標記的標籤。該模型被稱為雙向，因為在內部結合了兩個RNN：一個正向RNN從左到右處理序列，而一個反向RNN則從右到左處理序列。兩個RNN的輸出對於每個標記進行平均，以計算模型的最終標籤預測。

詞嵌入特徵

我們的模型以將目前的單詞映射到詞嵌入矩陣中的一列的索引作為輸入。該列是一個稱為詞嵌入的實值數值表示。詞嵌入可以被視為固定常數或可學習的參數。當固定時，假定詞嵌入矩陣的值已經經過預訓練，並且在訓練期間不進行更新。當可學習時，矩陣的值被視為模型參數，並且在訓練期間進行初始化，可以是預訓練的值或隨機初始化，並且在訓練過程中進行更新。我們對這三個選項進行了實驗。對於預訓練的詞嵌入，我們採用了一組已發表的、使用跳字模型算法在超過4億條推文上進行訓練的400維詞嵌入。這些詞嵌入的詞彙涵蓋了我們數據集中96.5％的標記。這些推文不特定於任何特定領域。此外，跳字模型算法是一種無監督機器學習方法，旨在編碼語義和句法信息，與ADR標記任務沒有直接關聯。這些詞嵌入是我們模型提供的唯一輸入特徵。雖然我們測試了其他特徵，但它們並未產生性能差異（請參見補充附錄）。

模型架構

我們的雙向長短時記憶遞歸神經網路模型為一個長度為 n 的推文中的每個詞嵌入產生一個ADR標籤，過程可以描述為在推文中的序列步驟 t = 0 到 t = n - 1（對應於推文中的詞位置）。在步驟 t = i 時，前向RNN為標記wi計算一個D維輸出，而後向RNN則為標記wn-1-i計算一個D維輸出。前向和後向鏈都是LSTM RNN，其隱藏層節點根據當前輸入（詞嵌入或前一隱藏層的輸出）和前一個輸入的個別輸出來計算其輸出。每個節點包含輸入、遺忘和輸出門，這些門加權當前輸入、前一隱藏狀態和當前隱藏狀態，以計算其當前輸出（見表1）。對於給定的標記wi，前向和後向鏈的最終隱藏層輸出進行元素-wise平均，並傳遞到一個完全連接的層，將D維向量減少為4個值，表示每個標籤的概率（<PAD>，I-ADR，I-Indication或O）（見圖2）。



圖二：我們雙向網絡架構的概觀，展示了對標記為wn-5¼addy的標記預測。該網絡以步驟方式處理整個長度為n的序列。在第ti步，前向RNN預測標記為wi的標記。在步驟tn-(iþ1)中，後向RNN預測標記為wi的標記。一旦處理完整個長度為n的序列，模型對每個標記的前向和後向預測進行平均。最終的全連接層將輸出維度減少為4，代表每個可能標記（<PAD>，I-ADR，I-Indication或O）的概率。

模型變種

我們評估了三種BLSTM變種：

* 方法1（BLSTM-M1）：單詞嵌入值以隨機初始化（標準正態分佈乘以我們詞彙量的平方根32）並視為可學習的模型參數。
* 方法2（BLSTM-M2）：單詞嵌入值以公開可用的預訓練數據集28,36初始化，並視為可學習的模型參數。
* 方法3（BLSTM-M3）：單詞嵌入值初始化方式與方法2相同，但視為固定常數。

我們使用Keras Python函式庫37和Theano38,39後端實現了這些模型。我們使用在線訓練（每次1條推文）進行訓練，並通過時間反向傳播40進行訓練，分別對BLSTM-M1、BLSTM-M2和BLSTM-M3模型進行了6、6和18個時代的訓練（即在訓練集上的遍數）。我們在訓練之前，使用訓練集上的交叉驗證校準了每個模型的時代數。在MacBook Pro上使用2.6 GHz處理器和16 GB RAM，將BLSTM-M3模型在完整的634條推文訓練集上進行18個時代的訓練花費了45分鐘。測試時間不到5秒。我們的模型代碼可在以下鏈接找到：[https://github.com/chop-dbhi/twitter-adr-blstm。](https://github.com/chop-dbhi/twitter-adr-blstm%E3%80%82)

基準序列標記方法

我們也考慮了兩種基準技術：詞庫匹配方法和CRF模型。

字典匹配方法

第一個基準方法通過匹配詞彙表中的概念來識別ADR提及。我們的詞彙表包含 13,014 個概念詞彙，結合了先前發表的ADR詞彙表21與來自Consumer Health Vocabulary41的概念，這些概念對應於已知與常用於治療ADHD的藥物相關的不良事件，該資訊來自國家兒童健康質量研究所的Vanderbilt Assessment Follow-up—PARENT informant表格。我們從概念詞彙創建了一個Apache Lucene索引，並在每個推文中查詢了概念匹配。我們對概念詞彙和推文標記進行了預處理，刪除了常見的英文單詞（例如，“the”，“of”），並將單詞縮短為其詞幹形式（例如，“jumped”變為“jump”）。我們認為當詞彙概念中的所有標記都存在於推文中時，匹配詞彙表視為一個預測的ADR，如果預測的ADR至少部分重疊於已標記的ADR範圍，則視為一個真陽性的近似匹配。我們的其他詞彙表詞彙和模型實現可在 <https://github.com/chop-dbhi/twitter-adrlexicon> 找到。

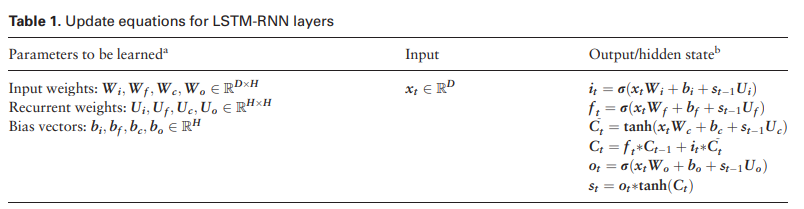
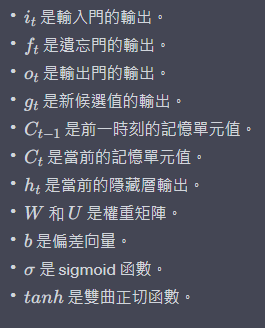


表1. LSTM-RNN 層的更新方程式



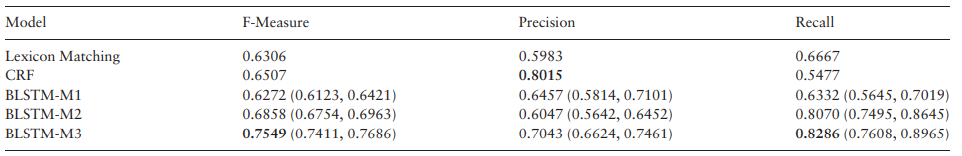
參數 D = 400 和 H = 256 分別表示詞嵌入和LSTM內部維度。參數 x\_t 表示序列中第 t 個標記的詞嵌入輸入到LSTM層。s\_t 表示當前輸入標記的當前層中節點的輸出，s\_t-1 表示前一標記的當前層中節點的輸出。輸入門 it 權重當前輸入的節點計算得到的隱藏狀態 C~t。遺忘門 ft 權重當前輸出計算中的節點的前一隱藏狀態 Ct-1。輸出門 ot 權重當前輸出，然後將其傳遞到下一層。內部激活函數 r 是線性近似的標準sigmoid，即 r(x) = max(0, min(1, 0.2 \* x + 0.5))。運算符 \* 表示逐元素乘法。

條件隨機場模型

我們的第二個基準方法使用了一個條件隨機場模型，如前述的最先進的研究所述。我們使用CRFSuite軟件42實施了該模型，並用以下特徵來表示每個標記：

* 上下文標記：當前標記及其之前3個標記和之後3個標記的標識。標記被轉換為小寫字母和數字，@-提及和URL被替換為特殊符號。
* ADR詞彙表匹配：二元特徵，指示當前標記是否與ADR詞彙表中的任何單詞匹配。21
* 詞性標記：當前標記的詞性標記。我們使用了ARK Twitter詞性標記器。34,43
* 否定：二元特徵，指示當前標記是否是否定的。我們使用DepND44-46進行否定檢測。
* 詞群集：與當前標記和上下文標記所對應的詞嵌入的最近詞群集。為了生成詞群集，我們使用了與BLSTM模型相同的預訓練詞嵌入28，並將其聚類為100組，使用k-means算法。我們還考慮了從與醫療保健相關的推文中導出的詞群集，但這些對F值幾乎沒有影響（請參見補充附錄）。

表2. 每个模型在10轮训练和评估中达到的近似匹配F1分数、精确度和召回率的平均值及95%置信区间。



粗體文字表示每個指標的最佳表現模型。

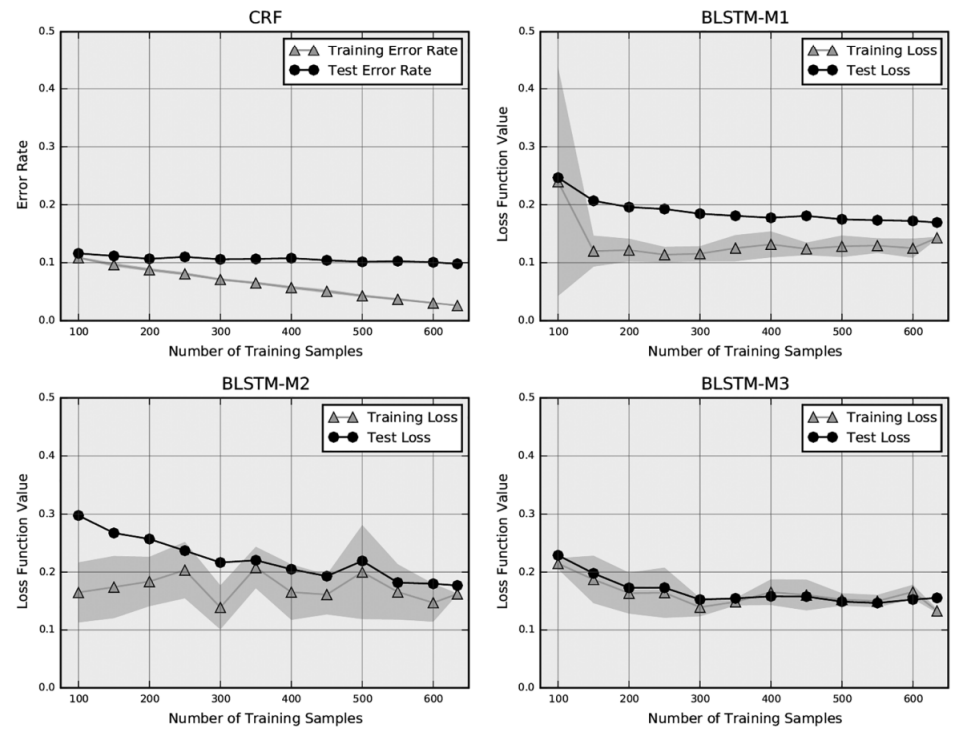


圖 3. 我們觀察每個模型在訓練集大小範圍從 50 到 641 個樣本的測試集上所達到的平均近似匹配 F1 分數（每個大小有 5 個隨機選取的訓練集）。

結果

我們使用近似匹配方法来评估模型的性能。近似匹配考虑如果预测到的ADR跨度与一个或多个实际ADR跨度重叠，则将预测到的ADR跨度视为正确。给定推文“The Seroquel gave me lasting sleep paralysis”，真实的ADR跨度是“sleep paralysis”，预测的跨度“lasting sleep paralysis”或简单的“paralysis”都将被视为正确。我们报告了近似匹配的精确度、召回率和F1分数： 精确度 = 正确识别的ADR跨度数 预测的ADR跨度数 召回率 = 正确识别的ADR跨度数 实际ADR跨度数 F1分数 = 2 \* 精确度 \* 召回率 精确度 + 召回率

我們模型在訓練期間處理推文的隨機順序可能會產生稍微不同的參數值，進而在留出的測試集上產生略有不同的預測結果。為了考慮這種隨機性，我們在完整的訓練集上對每個模型進行10次訓練並進行預測。我們報告了這10次訓練和評估過程中每個指標的平均值和95%信心區間。

各模型在包含210篇推文的測試集上的平均近似匹配結果見於表2。我們的BLSTM-M3模型在F1分數方面顯著優於所有其他模型（P < .01）（使用計算密集型的近似隨機化測試48–50計算的統計顯著性）。 雖然CRF獲得了最高的精確度，但BLSTM-M3模型在精確度和召回率的平衡方面表現更好，這反映在其F1分數中，因此具有新的最先進性能。初始化為預訓練嵌入的兩個BLSTM模型（BLSTM-M2和BLSTM-M3）的性能顯著優於基線和隨機初始化嵌入的BLSTM-M1模型（P < .01）。在詞彙匹配、CRF和BLSTM-M1模型的性能之間沒有顯著統計差異（P > .05）。

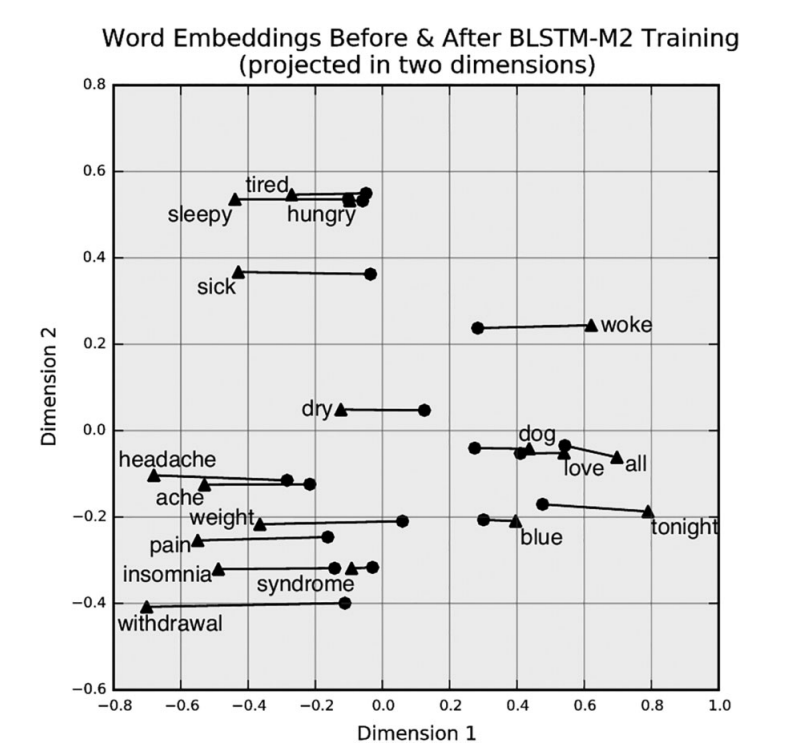


圖4. 使用奇異值分解投影到2維的BLSTM-M2模型訓練前（圓圈）和訓練後（三角形）的詞嵌入。 值得注意的是，在訓練後，與ADR提及相關的單詞（如sleep，pain，weight，hungry，dreams等）與非ADR相關的單詞（如blue，tonight，all，love等）分開。

討論

學習曲線分析

瞭解機器學習模型對數據樣本的敏感度（即模型變異性）非常重要，特別是為了確定樣本大小是否足夠。為了進一步研究這一點，我們檢查了圖3中顯示的模型學習曲線（隨著訓練集大小的變化而變化的性能）。對於BLSTM-M3模型，訓練和測試期間獲得的損失函數值接近於同一遞 asymptotic值。這通常表示變異性受到合理控制，因此增加更多的訓練示例不會改善模型性能。我們觀察到訓練和測試曲線的極限錯誤率為非零值（9％），這表明存在偏差，並且進一步的性能改進可能需要額外的輸入特徵。相反，CRF、BLSTM-M1和BLSTM-M2模型的訓練和測試錯誤率沒有達到相同的極限值。根據訓練和測試評估之間的錯誤率差異，CRF模型似乎受到變異性的影響最大。這可能是由於該模型需要更多的輸入特徵（即更多的自由度）。雖然對所有3個模型來說，更多的訓練示例可能會有所幫助，但學習曲線的下降斜率非常小，因此很可能需要過多的手動標記訓練示例來充分控制變異性。事實上，這正是使用預訓練詞嵌入的一個很有力的理由，因為它們可以使用無監督方法獲得，無需標記的數據集。

特徵分析

我們將 BLSTM-M3 模型的優異表現主要歸因於使用固定的詞嵌入，這些嵌入是專門設計來編碼語義信息，而與 ADR 檢測任務無直接相關。為了了解它們的影響，我們分別評估了模型的性能，相對於詞嵌入是否（1）隨機初始化或預訓練，或（2）在訓練期間保持固定或允許變化。

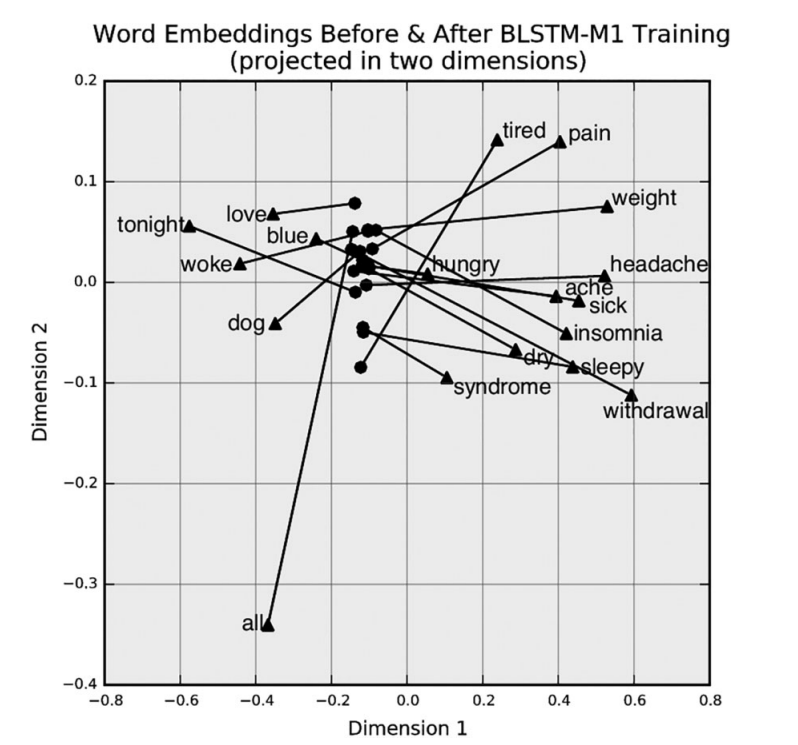


圖 5. 與圖 4 中相同單詞對應的嵌入，由我們隨機初始化詞嵌入模型 BLSTM-M1 學習而得。隨機初始化的模型同樣將與 ADR 相關和非 ADR 相關的單詞分開，但在語義上相關的單詞的嵌入不如預訓練的嵌入那麼密集（例如，“疲倦”和“昏昏欲睡”）。

我們首先考慮使用預訓練嵌入的影響，注意到使用預訓練嵌入的 BLSTM-M2 和 BLSTM-M3 模型的召回率分別比使用隨機初始化嵌入的 BLSTM-M1 模型高了 27.4% 和 30.8%。我們認為這是一個結果，原因在於對於預訓練嵌入，語義上相似的詞對在嵌入空間中幾乎有相等的嵌入，這隱含地讓模型瞭解測試集中的單詞，即使這些單詞在訓練集中未被觀察到。例如，考慮語義上相似的 ADR，如“疲倦”和“昏昏欲睡”。我們期望對應的詞向量在嵌入空間中接近。如果“疲倦”出現在訓練集中，而“昏昏欲睡”僅出現在測試集中，我們仍然可以期望模型檢測到測試集中的 ADR，因為對於“疲倦”激活的 RNN 節點也將激活對於“昏昏欲睡”的節點。相比之下，BLSTM-M1 模型隨機初始化嵌入，因此語義上相似的詞對可能相隔任意遠。這些詞對不一定會在訓練中收斂，因為語義相似性並不受 ADR 標記學習目標的加強，即使受到加強，測試集中不在訓練集中的單詞會被隨機向量表示。因此，在我們的例子中，模型永遠不會觀察到“昏昏欲睡”這個單詞，也不會將其識別為與“疲倦”相似。總的來說，這導致了隨機初始化模型的假陰性錯誤更多（召回率較低）。具體來說，我們在測試集中發現了 553 個不在訓練集中的獨特單詞，對應到 591 個測試集實例（84.6% 非 ADR，15.4% ADR），對於這些實例，BLSTM-M1、BLSTM-M2 和 BLSTMM3 模型的假陰性錯誤率分別為 13.0%、10.0% 和 7.6%。

接下來，我們考慮了相對於ADR任務更新詞嵌入的影響，注意到保持詞嵌入固定的BLSTM-M3模型，在相對於在訓練期間更新嵌入的BLSTM-M2和BLSTM-M1模型，實現了分別高達16.5％和9.1％的更高精度。我們假設這是因為BLSTM-M1和BLSTM-M2模型在詞嵌入中直接編碼ADR檢測信息。這在圖4和圖5中以定性方式進行了說明，我們觀察到ADR相關和非ADR相關的單詞在嵌入空間中分離。這種初始區別是死板的，並且會導致RNN偏向於將某些單詞解釋為ADR相關或非ADR相關，而不考慮其上下文。因此，在訓練期間見過的ADR相關單詞在測試時會傾向於被視為ADR相關，而不考慮上下文，並導致偽陽性錯誤（降低了精度）。這一假設得到了表3中數據的支持，我們在該表中對每個模型造成的偽陽性錯誤進行了表徵。對於固定嵌入模型（BLSTM-M3），偽陽性錯誤的10％歸因於ADR相關單詞，即在訓練中真實ADR範圍中出現的單詞在測試時被錯誤標記為I-ADR的情況。對於變嵌入模型，ADR相關單詞錯誤的比例要高出3倍。

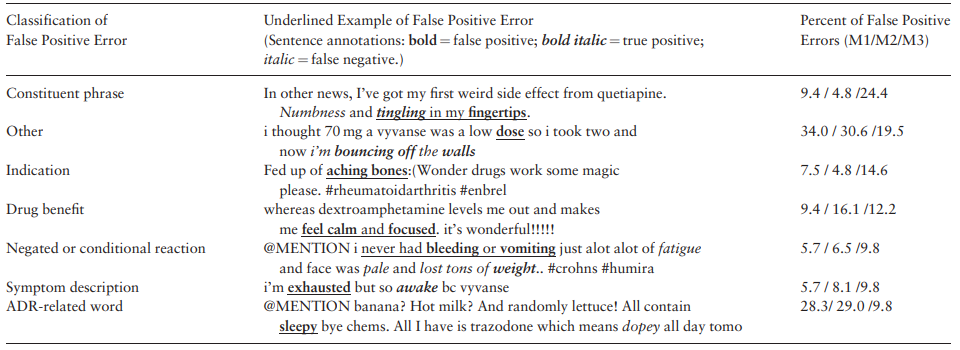


表3. 作者對BLSTM模型造成的偽陽性錯誤的分類

以下是各錯誤類型的例子，第二欄是底線標示。其他句子標註如下：偽陽性ADR範圍以粗體標示，真實陽性範圍以粗斜體標示，偽陰性範圍以斜體標示。

錯誤分析

我們進一步定性地檢視了BLSTM-M3模型的錯誤，尤其是偽陽性，如表3所示，以找出可能的原因。將近25%的偽陽性可以歸類為構成片語的錯誤，這意味著模型標註了屬於真實ADR的構成片語中的詞，但與註釋範圍不相鄰。這些錯誤通常對ADR進行了詳細說明，就像表3中的第一個例子，我們的模型將單詞“fingertips”標識為“tingling in my fingertips”這個構成片語的一部分，但嚴格的ADR邊界只包括動詞“tingling”。其他常見的錯誤包括將指示、藥物效益、ADR症狀、否定反應和其他與ADR相關的詞標記為ADR。

限制與局限性

研究的潛在限制之一是我們的Twitter數據集的大小和範圍。具體來說，有限的數據規模可能會限制我們將結果概括到可能在Twitter和其他社交媒體平台上提及的所有藥物和ADR類型。可能存在很少提及的ADR，或者與我們的搜索詞不匹配的藥物相關的ADR，這些ADR可能具有在我們的訓練集中未呈現的語義和句法模式。同樣，來自其他數據來源的文本可能包含不同的語言模式，需要創建額外的詞嵌入。此處使用的預訓練詞嵌入嵌入了與Twitter上使用的語言相關的語義信息。為了在具有顯著不同語言的數據語料庫上正確評估我們的模型，例如臨床筆記，需要創建新的詞向量。幸運的是，相關的數據集已經存在（例如，多參數智能監護在重症監護[MIMIC] III），這是我們持續研究的重點。

結論

與其他自然語言處理任務的近期成功相呼應，我們的研究結果顯示，具有預先訓練詞嵌入的循環神經網路模型能夠有效地識別社交媒體數據中的不良反應，特別是 Twitter 數據，並通過達到在 F1 得分方面具有統計學上顯著的優越表現，建立了新的最先進性能，超越基於 CRF 模型的性能

固定詞嵌入中所編碼的語義信息來自於大型、非醫療特定的數據集，是改善模型性能的關鍵因素。特別是，這些語義信息通過讓模型能夠在測試集中識別在訓練中未觀察到的不良反應，從而提高了模型的召回率。我們還注意到，模型訓練所需的ADR標記示例較少，就能達到最佳性能。這是有利的，因為它減少了所需的ADR標記訓練示例的數量，而這一點很重要，因為標記是一個資源密集型的過程。此外，詞嵌入可以使用無監督學習方法創建，只需要未標記的數據，因此可以輕鬆擴展到大型數據集。最後，這些詞嵌入是我們模型使用的唯一特徵，因此避免了特定任務的特徵工程，這表明了對額外的序列標記任務具有泛化能力，並且相對於需要專家設計的昂貴且難以開發的輸入特徵的系統，具有簡單性。

雖然我們的結果令人鼓舞，但標記單詞的任務只是與自動ADR識別相關的挑戰之一。在我們的研究中，模型主要是在包含與藥物相關事件的用戶帖子上進行訓練和評估的。我們發現，將許多不包含ADR的帖子納入訓練數據，對於RNN和基線模型都產生了相對較差的結果。因此，一個完全自動化的系統可能需要兩個額外的組件：（1）高度準確的二元分類器，可預測文本是否包含ADR，以及（2）一種通過文本標記過程識別出的類似ADR（例如，胃疼和胃痛）的方法，將其映射到單一本體術語，以支持聚合分析。對於前者的方法在先前的研究中已有描述。至於後者，在藥物警戒方面受到的關注較少。鑒於我們的深度學習架構的效益，再加上從大型未標記數據集中創建的詞嵌入中編碼的語義信息對ADR檢測的重要性，我們將在未來的工作中探索相關方法，將識別出的ADR文本映射到本體術語的適用性。

備註:

ADR: "ADR" 代表 "Adverse Drug Reaction"，中文翻譯為「不良藥物反應」。這是指使用藥物時可能會引起的不利效應或負面反應，可能對患者的健康造成損害或不良影響。在藥物研究和監測中，識別和監控不良藥物反應對確保藥物的安全性和有效性至關重要。