遞迴式類神經網路在語言模式處理上的研究

The Study Of Recurrent Neural Networks For Language Modeling

王文俊，李俊曉，劉繼謚
中華電信股份有限公司 中華電信研究所

摘要

在語音辨認及音韻分析處理，遞迴式類神經網路(recurrent neural network)已被廣泛利用並得到不錯的結果，它的特性是將前一時段的輸出層或隱藏層的輸出以延遲的方式與下一時段之輸入訊號一併輸入網路進行學習，因此可以有效利用時間軸上的資訊。本文則嘗試利用此種類神經網路架構進行語言模式的處理，實驗的語料是採用漢語平衡語料庫，而由於考慮參數量及複雜度，因此所有訓練及測試都是直接在詞類串列上進行。實驗的進行是以目前詞之詞類作為輸入，而以下一詞之詞類為目標值，對網路的各項連接加權值作最佳化的調整。對於此種方法所得結果的評估，一方面是和詞類雙連文法(part-of-speech bigram) 及詞類三連文法(part-of-speech trigram)比較，觀察其相似程度；另一方面則是觀察真實詞類串列和神經網路預測值之差異藉以判斷其能否協助決定子句段落的位置。實驗結果顯示利用類神經網路可以極類似於雙連文法及三連文法，除此以外此方法也有助於文句分析以決定子句段落位置。

一、 前言

文字處理是語音合成的重要程序，除了正確的字轉音必須在此完成以外，子句段落位置的決定也是一項重要工作。而由於大部份語音合成系統都希望能達到即時處理的要求，以及精確的文句分析絕對需要具備深厚的語言學基礎，使得這項工作並不容易完成。本篇論文則嘗試利用遞迴式類神經網路架構進行此項研究。遞迴式類神經網路已被廣泛利用在語音辨認[Hunt 1993]及音韻分析處理上，並得到不錯的結果。它的特性是將前一時段的輸出層或隱藏層的輸出以延遲的方式與下一時段之輸入訊號一併輸入網路進行學習，因此可以有效利用時間軸上的資訊。而文句中的文字串列或詞類串列也同樣具有時軸上前後的關係，這樣的關係能否用遞迴式類神經網路進行學習是本篇論文的探討重點。實驗將對漢語平衡語料庫的文句資料進行學習，根據大部份統計模式的原理，在多次的累進訓練後，高出現頻率的詞類串列必然會得到較低的誤差值。而假設網路的學習效果很好的話，則在一些子句的轉接處，神經網路的預測值和目標值的差異將會變大，而這樣的變化即可被運用來決定子句段落位置。

以中文常用字 5401 個字及常用詞 80000 個詞進行分析，將會有訓練語料不足、參數量太大等問題，因此在此將只探討詞類間的關係。也就是利用已作好詞類標示的漢語平衡語料庫，將每一個詞的詞類依序作為輸入，而以下一個詞的詞類為目標值，對網路的各項連接加權值作最佳化的調整。評估實驗結果時，除了與實際值作比較外，訓練結果和詞類雙連文法(part-of-speech bigram) 及詞類三連文法(part-of-speech trigram)的相似程度也是比較的重點；另一方面則也將觀察真實詞類串列和神經網路預測值之差異藉以判斷其能否協助決定子句段落的位置。

二、 遞迴式類神經網路

遞迴式類神經網路在很多方面都和傳統的前進式神經網路(feed-forward neural network)類似，最大的不同點是在它的架構變化如圖一所示，將隱藏層或輸出層經過延遲後和下一時段之輸入特徵一併再作輸入，其餘小變化的作法也包括將延遲的數目增加以及在不同層加入不同的輸入等等。遞迴式類神經網路對時軸序列訊號的分類有非常不錯的效果，它可以學習到訊號間的複雜關係，因此對大多數語音辨認處理而言，遞迴式類神經網路是一種非常好的架構。

至於利用遞迴式類神經網路在語言模式應用的例子[Elman 1990]則有屬於小範圍的文字預測，這類實驗是利用約 30 至 40 個詞構成一些短句子，再將這些句子輸入神經網路進行訓練，訓練過程是以下一個詞為目標值，而隱藏層的輸出則可以被用來作類似詞類的分類，因為隱藏層的輸出呈現出有限狀態機器(finite state machine)的分佈，而這些有限狀態是受到前後詞的影響。至於本文的運用則是以經過詞類標示之語料進行訓練，輸入目前詞的詞類而將目標值定為下一個詞之詞類。

如上所述，遞迴式類神經網路的優點是對訊號具有分類的能力，不過卻需要耗費相當大的計算資源，雖然有很多方法被提出來提高訓練速度及分類的準確性，如加權值調整的方法，輸出轉換函數的選擇以及輸出延遲等，但本文將不在這些地方進行討論，而將探討此種網路架構是否有助於解決本文所提出的問題以及增加延遲的數目如圖二所示能否使模式更精準，另外為避免訓練耗時因此訓練語料並未使用整個漢語平衡語料庫，而有關訓練語料庫大小對實驗結果的影響也將在此作比較。

三、實驗描述

本文之實驗設計除想比較神經網路與雙連文法或三連文法之差異，另外也想探討資料量對文法模式訓練之影響。因此希望能比較出語料庫大小的影響，目前是用漢語平衡語料庫之前五十個檔案共約 140 萬詞構成一個大語料庫，而小語料庫的設計則為考慮文句涵蓋範圍而將漢語平衡語料庫之各個檔案隨機選取數句構成一個約含三萬詞的測試語料庫。

實驗的進行是利用上述之大小語料庫分別去統計出詞類雙連文法及詞類三連文法，另外以小語料庫進行遞迴式神經網路之訓練，訓練時是輸入目前詞之詞類，而以下一詞之詞類作為理想值，網路之架構為 55 個輸入節點(46 類 POS 及 9 類標點符號[中研院 1995])，60 個隱藏層節點，55 個輸出層節點。為加快訓練速度及收斂速度，我們同時採用了累進訓練(incremental training)的方法，也就是說在頭幾次的訓練僅使用到少量短句的語料，待接近收斂後再增加語料。如此的作法即可使訓練速度加快，訓練之結果隨著次數增加其誤差值呈現遞減的變化。有關最佳一至五的結果含有理想值之包含率如表一所示，訓練結果的好壞若直接與理想值作平均誤差之估算其結果實不令人滿意，因此我們也將評斷此結果與雙連文法及三連文法的接近程度，此結果則如表二所示，同時我們也將訓練結果和一個無機率之模式比較，也就是所有詞類均是任一詞類之可能後接詞類且其機率值均相同。而多延遲遞迴式類神經網路能否使預測模式與雙連文法、三連文法甚至於 N-gram 文法更加接近也有初步的結果將在下節介紹。

四、 實驗結果與討論

由表一可看出訓練結果的 top 1 正確率並不高，但是此模式的目標原本就不是在將預測正確率調整至百分之百，而是希望能儘可能縮小和其他統計式機率模式的差異。由表二可看出訓練結果和雙連文法非常接近，而和三連文法差異稍大；因此我們原本寄望多延遲的架構能模擬成更多詞的影響，而能降低與三連文法的差異，但由表三及表四的比較，多延遲模式所得之結果並未如預期比單一延遲之效果好，這似乎說明因為中文的特性，對詞的定義不夠明確，有些長詞都可能被斷成數個短詞，以致於 2 個或 3 個的延遲並不直接和 trigram 及 4-gram 等效。另外由表中的數據也可以看出我們提供了分別由大小語料庫統計出的雙連文法及三連文法，而發現訓練結果與得自大語料庫的統計結果較接近。

雖然利用此方法要作到自然語言了解的地步還有很大的距離，但在提供一個迅速且有效的子句段落預測模式仍有一些效果。如圖三及四所示之測試句子，在平均誤差曲線的各個谷底處都有較大可能被視為一個子句段落位置。而在審視過大部份的測試語句也發現在連接詞及副詞之前都會有明顯變化，而對於“中”、“裡”等後置詞可能是因為出現次數過低使得變化並不一致，另外在“的”之後若接名詞則會在該名詞之後產生變化，但若接具名物化之述詞時，因為在詞類選擇時將包括名物化的所有特徵都捨棄，以致於在此狀況時，大部份的變化都落在“的”之後。總括而言，對於本文所提出的模式仍有許多地方亟待改進，包括詞類分類的選擇，訓練語料的擴大以及神經網路的架構調整，而一些語法上的特殊修飾也是必須作詳細討論。

五、 參考文獻

1. Andrew Hunt, “Recurrent Neural Networks for Syllabification”, Speech Communication, vol. 13, pp. 323-332, 1993.

2. Jeffrey L. Elman, “Finding Structure in Time”, Cognitive Science, vol. 14, pp. 179-211, 1990.

3. 中央研究院，詞庫小組，“中央研究院平衡語料庫的內容與說明”， Technical Report no.95-02, 1995.

表一. 前一及前五含最佳結果之包含率

	top 1	top 5	top 10
inclusion rate	32%	63%	78%

表二. 訓練結果與得自不同語料庫之文法的平均誤差

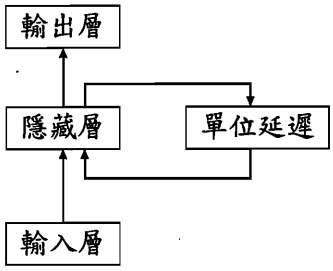
	unigram	bigram	trigram
small corpus	0.40156	0.19432	0.28523
large corpus	0.37294	0.15756	0.24137

表三. 多延遲與否的訓練結果與得自小語料庫文法的平均誤差

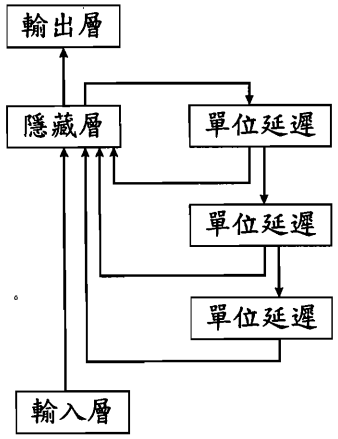
small corpus	bigram	trigram
without md	0.19432	0.28523
with md	0.19733	0.29058

表四. 多延遲與否的訓練結果與得自大語料庫文法的平均誤差

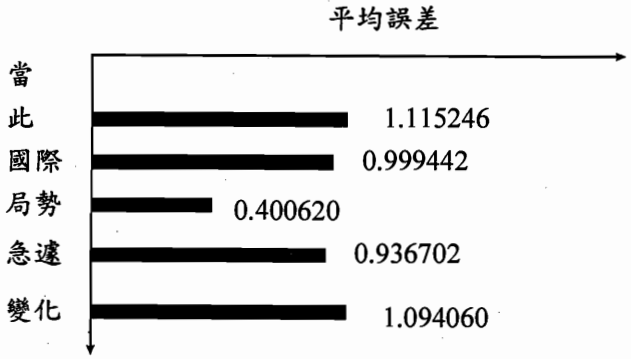
large corpus	bigram	trigram
without md	0.15756	0.24137
with md	0.16276	0.24542



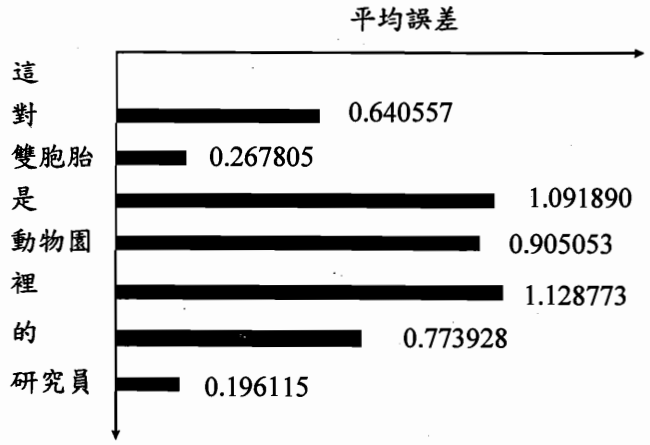
圖一. 遞迴式類神經網路之基本架構



圖二. 多延遲遞迴式類神經網路之架構



圖三. 測試句子1 之平均誤差結果



圖四. 測試句子2 之平均誤差結果