Lecture 5 : Sequence Tagging

楊立偉教授
台灣大學工管系

wyang@ntu.edu.tw

Class Topics We Have

- L1: Document Similarity
  - "a Bag of Words" Model, Term Weighting (TF-IDF, $\chi 2$, MI)
  - Vector Space Model (VSM)
- L2: Co-occurrence
  - Association
  - Link analysis : Co-citation & Coupling
- L3: Classification
  - Naïve Bayes, k Nearest Neighbors, Support Vector Machine
  - Decision Tree, Bagging and Boosting, Random Forest
  - Neural Networks

- L4: Clustering
  - k Means, DBSCAN, Hierarchical Agglomerative Clustering
  - Topic Modeling
- L5: Sequence Tagging
  - Language model, HMM, CRF, RNN
- L6: Language Processing
  - Chinese Processing and other language issues
  - Word Embedding : LSA, Word2Vec, GloVe
  - Word window classification

Sequence Tagging

- To assign of a label to each member of a sequence of observed values. For example:
  - part of speech tagging and voice recognition in language processing 語意分析中的詞性標記及語音辨識
  - Applications of sequence analysis or prediction in finance / bioinformatics 財務或生物資訊上的序列分析應用
- This can be done…
  - as ***a set of independent classification tasks***. For example, step forward one per member of the sequence a time.
  - by making the optimal label for a given element dependent on the choices of ***nearby*** elements.

Predict the next element ?

- Try to predict the next element in a sequence of data

current (and previous ones)	candidates of the next one
馬	(馬)上、(馬)虎、(馬)腳、(馬)達、(馬)屁、(馬)克…
馬 英	(馬英)九

wyang@ntu.edu.tw

Example (1)

- "美國是個自由的國家"
  - moving a cursor from left to right, use the last n words to tag the current one.
  - as a classification task, for example, to Decision Tree or SVM.

W_{i-4}	W_{i-3}	W_{i-2}	W_{i-1}	W_i
美	國	是	個	自
國	是	個	自	由
是	個	自	由	的
個	自	由	的	國
自	由	的	國	家

Use the last 4 words to tag the current one

- "美國是個自由的國家"

W_{i-4}	W_{i-3}	W_{i-2}	W_{i-1}	W_i
美	國	是	個	自
國	是	個	自	由
是	個	自	由	的
個	自	由	的	國
自	由	的	國	家

forward predicting ↑

W_{i+1}	W_{i+2}	W_{i+3}	W_{i+4}	W_i
國	是	個	自	美
是	個	自	由	國
個	自	由	的	是
自	由	的	國	個
由	的	國	家	自

↑ backward predicting

W_{i-2}	W_{i-1}	W_i	W_{i+1}	W_{i+2}
美	國	是	個	自
國	是	個	自	由
是	個	自	由	的
個	自	由	的	國
自	由	的	國	家

← bidirectional predicting

Example (1)

- "美國是個自由的國家"
  - to obtain probability $P(W_i|W_{i-1}, W_{i-2},\dots, W_{i-n})$
  - for the sequence, when n=1, the probability is

P("美國是個自由的國家")=

P(國|美)xP(是|國)x P(個|是)xP(自|個)x P(由|自)xP(的|由)x P(國|的)xP(家|國)

* instead of product of the probabilities, sum of log is used usually in programming.

Example (1)

- "美國是個自由的國家"
  - in voice recognition (語音識別) and some intelligent input methods (輸入法), the probabilities of various candidates are evaluated, and the highest one is chosen.

P(“美國是個自由的國家”)=0.08 ←chosen, for example

P(“美國是個製油的國家")=0.07

P(“美國似個自由的國家”)=0.05

P(“美國事個自遊的國家")=0.04

Variant (1)

- Larger size of the sliding window

for window size=2

P("美國是個自由的國家")=

P(是 | W_{i-1}=國∩ W_{i-2}=美) x P(個 | W_{i-1}=是∩ W_{i-2}=國) x P(自 | W_{i-1}=個∩ W_{i-2}=是) x
P(由 | W_{i-1}=自∩ W_{i-2}=個) x P(的 | W_{i-1}=由∩ W_{i-2}=自) x P(國 | W_{i-1}=的∩ W_{i-2}=由) x
P(家 | W_{i-1}=國∩ W_{i-2}=的)

可經由大量語料做次數統計而得

Variant (2)

- Word/n-gram instead of Character

P("美國是個自由的國家")=
P(國是|美國) x P(是個|國是) x P(個自|是個) x P(自由|個自) x P(由的|自由) x P(的國|由的) x P(國家|的國)

Sliding window size = 2 與 2-gram 哪個方法好？

→ 2-gram效果類似猜連續2字，組合數(類別)變多，需要的訓練資料要足夠

Test the language model in your brain

- use the surrounding ***unigrams*** to predict

這口天口氣口冷口，所口都口到口較口才口門， 不口夠口間口早口。 每個空格都有若干候選字

- use the surrounding ***bigrams*** to predict

這幾口天氣口冷了，所以口睡到口較晚口出門， 不太口時間口早餐。 大幅減少空格中的候選字數

- one of the possible answers

這幾天天氣變冷了，所以都睡到比較晚才出門， 不太夠時間吃早餐。

Test the language model in your brain

- the window vs. the order

最近的研究表示，漢字序順並不定一影閱響讀。

- for auto-correction : can you read this ?

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the human mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

From : http://www.mrc-cbu.cam.ac.uk/people/matt.davis/Cmabrigde/

Test the language model in your brain

- ***"You shall know a word by the company it keeps"*** (Firth, J. R. 1957:11)
- answer what "令和" is, after reading the following :

> 昨天日本公布新年號「令和」後，掀起一股新風潮。商家開始大搶「令和」商機，還有店家祭出，只要名字中有其中一個字，就有折價。
> 日本民眾：「咦？是什麼？令和？！令和，令和元年。」
> 日本學生：「新年號，令和。」

for example

2-gram	occurrences in bi-dir. window of 5
年號	2
商機	1
元年	1

wyang@ntu.edu.tw

Example (2) for financial applications

- 若想利用股票今日價量尋找明日價格漲跌之關係，採用序列模型編碼如下

	Day_1	Day_2	Day_3	Day_4	Day_5	Day_6	Day_7	Day_8
價格	升	升	平	升	降	降	降	平
成交量	升	升	降	降	升	降	平	平

- 同樣可以使用分類、或連續的條件機率進行分析

Example (2)

- 使用分類演算法來預測
  - 假設n=2，編碼方式為$D_{i-2價}$、$D_{i-2量}$、$D_{i-1價}$、$D_{i-1量}$，預測$D_{i價}$

	Day_1	Day_2	Day_3	Day_4	Day_5	Day_6	Day_7	Day_8	Day_9
價格	升	升	平	升	降	降	降	平	?
成交量	升	升	降	降	升	降	平	平	

i	$D_{i-2價}$	$D_{i-2量}$	$D_{i-1價}$	$D_{i-1量}$	$D_{i價}$
3	升	升	升	升	平
4	升	升	平	降	升
5	平	降	升	降	降
6	升	降	降	升	降
7	降	升	降	降	降
8	降	降	降	平	平
9	降	平	平	平	?

forward predicting,
for example

Example (2)

- 使用連續的條件機率來預測
  - 假設n=2，編碼方式為$D_{i-2價}$、$D_{i-2量}$、$D_{i-1價}$、$D_{i-1量}$，預測$D_{i價}$

	Day_1	Day_2	Day_3	Day_4	Day_5	Day_6	Day_7	Day_8	Day_9
價格	升	升	平	升	降	降	降	平	?
成交量	升	升	降	降	升	降	平	平	

$P(D_{i價}=升 \mid D_{i-1價}=平 \cap D_{i-1量}=平 \cap D_{i-2價}=降 \cap D_{i-2量}=平)=?$

$P(D_{i價}=平 \mid D_{i-1價}=平 \cap D_{i-1量}=平 \cap D_{i-2價}=降 \cap D_{i-2量}=平)=?$ 三者取機率大者為猜測

$P(D_{i價}=降 \mid D_{i-1價}=平 \cap D_{i-1量}=平 \cap D_{i-2價}=降 \cap D_{i-2量}=平)=?$

將條件機率各項展開成可算之項後求(近似)解，比較大小

Hidden Markov Model

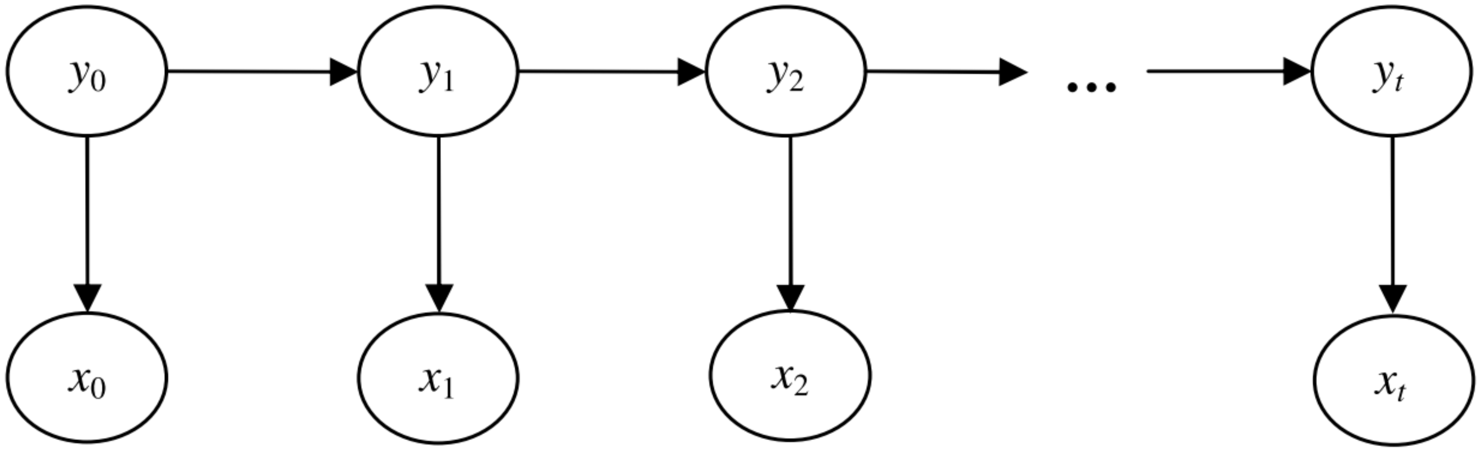


Fig. 3 Hidden Markov model

We have

$$Y = \langle y_0, y_1, \ldots y_t \rangle = \text{hidden state sequence}$$

$$X = \langle x_0, x_1, \ldots x_t \rangle = \text{observation sequence}$$

Ref. Lei Zhang and Bing Liu, "Aspect and Entity Extraction for Opinion Mining", 2014

HMM models a sequence of observations X by assuming that there is a *hidden* sequence of states Y. Observations are dependent on states. Each state has a probability distribution over the possible observations. To model the joint distribution $p(y, x)$ tractably, two independence assumptions are made. First, it assumes that state y_t only depends on its immediate predecessor state y_{t-1}. y_t is independent of all its ancestor $y_1, y_2, y_3, \ldots, y_{t-2}$. This is also called the *Markov* property. Second, the observation x_t only depends on the current state y_t. With these assumptions, we can specify HMM using three probability distributions: $p(y_0)$ over initial state, state transition distribution $p(y_t \mid y_{t-1})$ and observation distribution $p(x_t \mid y_t)$. That is, the joint probability of a state sequence Y and an observation sequence X factorizes as follows.

$$p(Y, X) = \prod_{t=1}^{t} p(y_t \mid y_{t-1}) p(x_t \mid y_t) \tag{2}$$

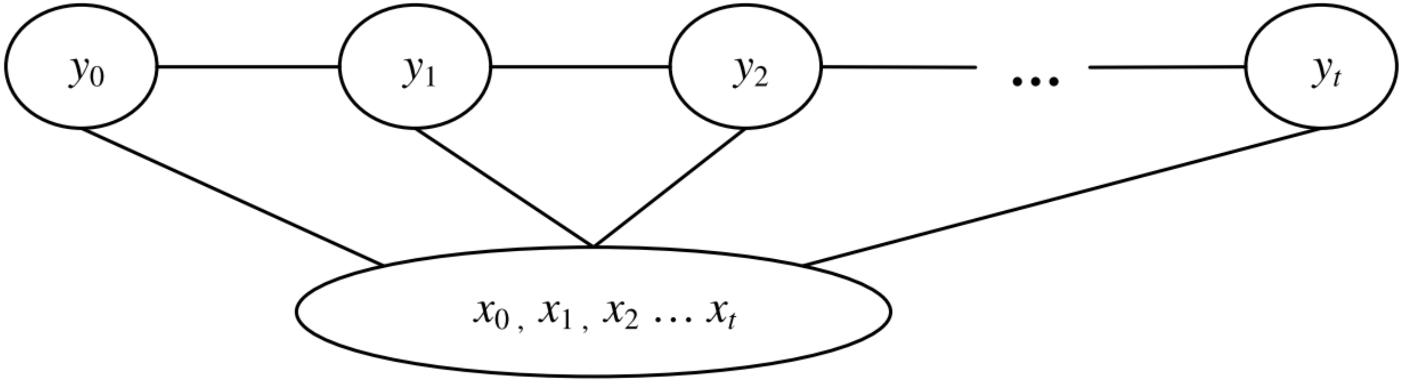
where we write the initial state distribution $p(y_1)$ as $p(y_1|y_0)$.

Given some observation sequences, we can learn the model parameter of HMM that maximizes the observation probability. That is, the learning of HMM can be done by building a model to best fit the training data. With the learned model, we can find an optimal state sequence for new observation sequences.

Ref. Lei Zhang and Bing Liu, "Aspect and Entity Extraction for Opinion Mining", 2014

Conditional Random Fields

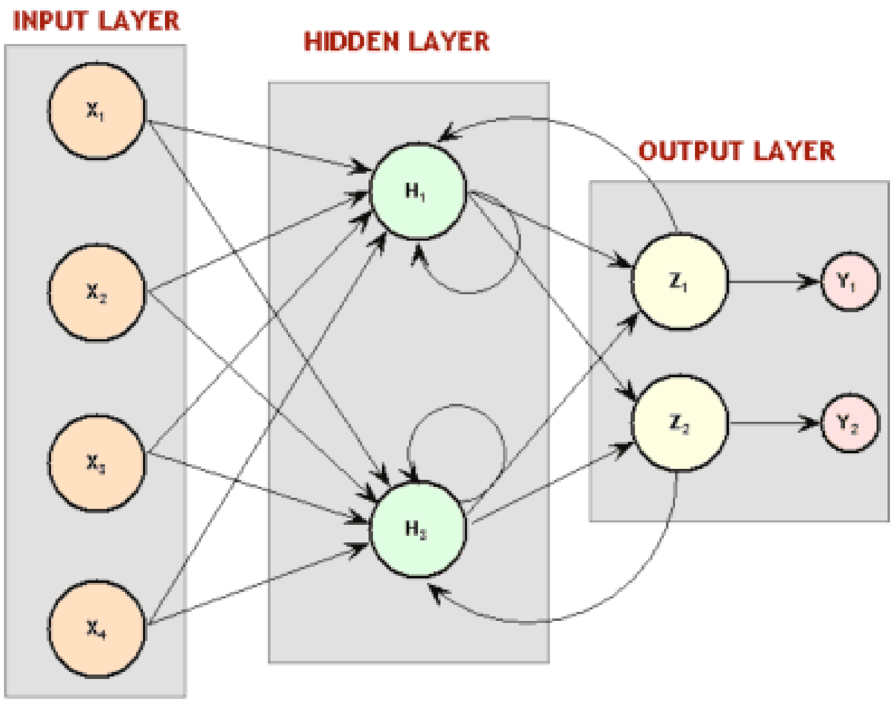
One limitation of HMM is that its assumptions may not be adequate for real-life problems, which leads to reduced performance. To address the limitation, linear-chain Conditional Random fields (CRF) (Lafferty et al., 2001; Sutton and McCallum, 2006) is proposed as an undirected sequence model, which models a conditional probability $p(Y|X)$ over hidden sequence Y given observation sequence X. That is, the conditional model is trained to label an unknown observation sequence X by selecting the hidden sequence Y which maximizes $p(Y|X)$. Thereby, the model allows relaxation of the strong independence assumptions made by HMM. The linear-chain CRF model is illustrated in Figure 4.



Ref. Lei Zhang and Bing Liu, "Aspect and Entity Extraction for Opinion Mining", 2014

- CRF implementation
  - CRFsuite http://www.chokkan.org/software/crfsuite/
  - CRF++ https://taku910.github.io/crfpp/
  - MALLET http://mallet.cs.umass.edu/

- Recurrent Neural Networks (RNN)



Ref: Recurrent neural network based language model. 2010

- Long Short-Term Memory (LSTM) RNN

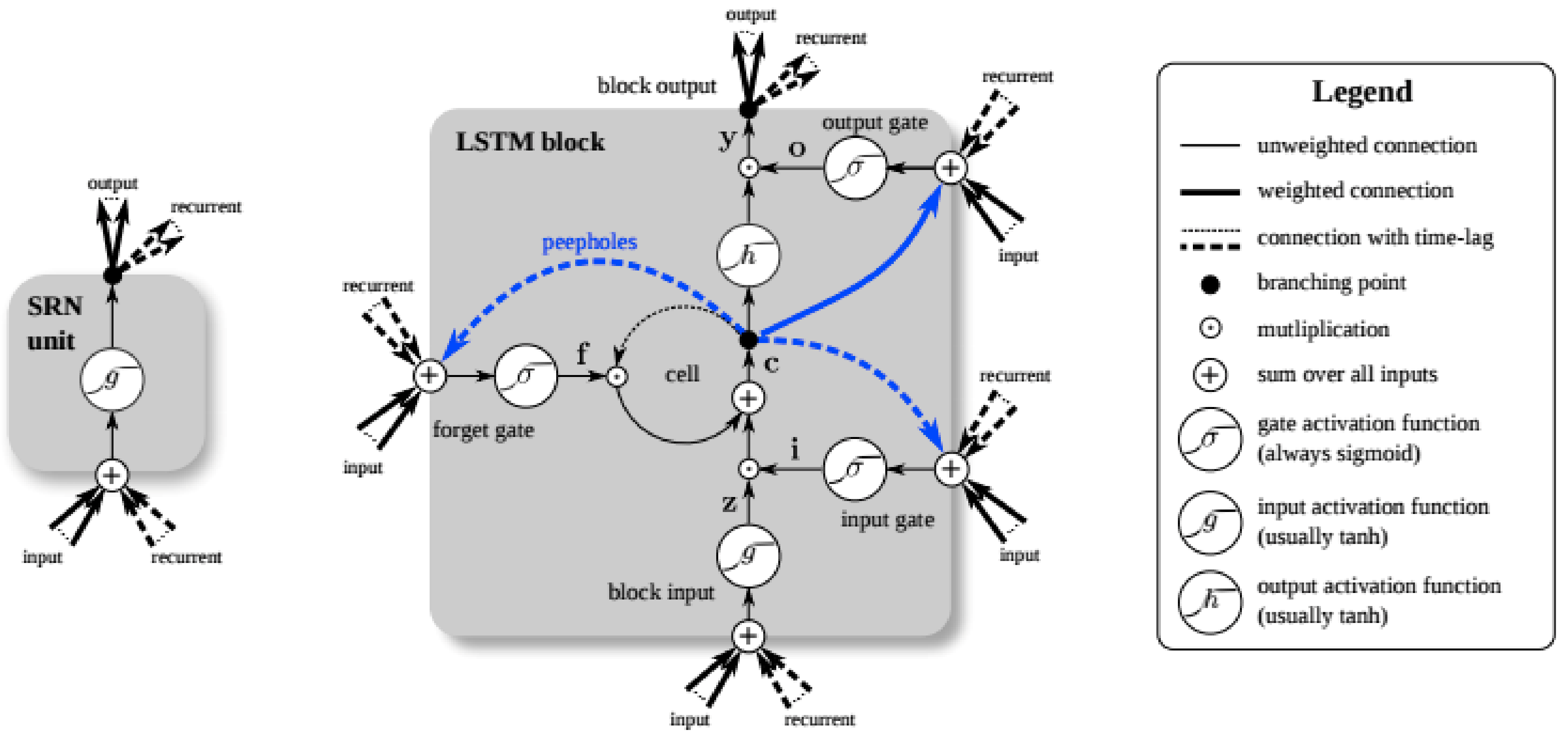


Figure 1. Detailed schematic of the Simple Recurrent Network (SRN) unit (left) and a Long Short-Term Memory block (right) as used in the hidden layers of a recurrent neural network.

Ref: http://193.6.4.39/~czap/letoltes/IS14/IS2014/PDF/AUTHOR/IS141304.PDF, 2014

Comparison : ARIMA

- ARIMA model (Autoregressive Integrated Moving Average model) is well-known time series analysis approach invented in 1970's.
  - 透過數列的自迴歸進行預測，利用移動平均項數及差分來穩定數列，減少波動；大量應用在金融及經濟學領域
- 思考 1 : ARIMA與本次介紹兩種方法的優劣比較？
- 思考 2 : 若是非同期對應的序列資料，該如何處理？

Discussions