陸、計算語言學中常用的公式

一、互見訊息(mutual information):

互見訊息(mutual information)是訊息理論(information theory)中的基本概念，計算的方式是兩個事件共同出現的機率除以個別事件出現的機率的積除再取以二為底的對數。

$$MI(x,y)=\log_2 \frac{P(x,y)}{P(x)P(y)}$$

如果只考慮緊鄰的兩個詞，則可代入下列公式。其中 N 代表總詞數，f(x,y) 代表 x 與 y 一起出現的次數 f(x)，f(y)分別代表 x 出現的次數與 y 出現的次數。

$$MI(x,y)=\log_2 \frac{\frac{f(x,y)}{(N-1)}}{\frac{f(x)}{N}\times\frac{f(y)}{N}}\cong\log_2 \frac{N\times f(x,y)}{f(x)f(y)}$$

Ken Church (1991)與他的同事率先提出以互見訊息計算詞與詞之間的相連性(word association)[5]。互見訊息值越高表示詞的相連性越高，當語料庫夠大時，而互見訊息值大於零，表示這兩個詞常常一起出現很可能是搭配語(collocations)，成語，或常見的人名，地名。利用互見訊息可以從中文語料庫中自動抽取詞彙。

二、T-值(t-score):

互見訊息可以視為一種相似度測量，T-值則可以視為相異度的測量。T-值

[5] Using Statistics In Lexical Acquisition.In Zernik, U. (1991) (eds.) Lexical Acquisition: Exploiting On-Line Resources to Build a Lexion.

(t-score)是計算語言學中常用的統計顯著性的檢定(statistical significance test)，也是 Ken Church (1991)與他的同事率先運用在計算語言學，通常與互見訊息搭配一起使用。T-值與標準差和信賴區間(confidence interval)密切相關。當語料庫夠大時，而 T-值大於 1.65 時表示有 95%的信心證明差異是存在。

T 值的計算可以採用下列簡化的公式。

$$t \approx \frac{f(x,y) - \frac{f(x)f(y)}{N}}{\sqrt{f(x,y)}}$$

三、熵(entropy):

熵(entropy)也是訊息理論中的基本概念，1960 年代被用於資訊檢索，1990 後被廣泛運用到計算語言學中。熵可視為是測量驚奇，不確定性，或訊息的量。公式如下。

entropy = - k Σ pi log(pi)

四、N 連詞(ngram)語言模型(language model)是是統計式計算語言學中最簡單也最常用的語言模型。N 連詞語言模型假設一個句子第 N 個詞的機率可以由前面的 N-1 個詞決定。雖然這個假設過於簡化語言的複雜性，但在語音辨識與其它應用上不失為一個有效的方法。最常用的是二連詞(bigram)與三連詞(trigram)模型。