因為一個有 l 個詞的句子,每個詞最多只會修飾另一個詞,所以整個句子最多只有 l 個關係。也因為如此, lx(l-l) 個詞組中大部分都是沒有關係的(0), 所以計算正確的預測兩詞關係很容易達到很高的正確率。故這邊不討論預測兩詞關係的正確率,而討論有多少詞預測修飾的對象是正確的。結果參見下表:

表（十）

正確預測修飾對象詞	57109（76.298%）
結構完全正確的句子數	6724（53.826%）

拾貳、多義詞詞義辯識

一個詞可能有好幾個不同的意思,例如 bank 有銀行,河堤,庫等多個意義。詞義辨識的目的就是要讓電腦自動辨識一個岐義詞在某一個語境裡正確的意義。由於現有詞性標記的演算法正確率都相當的高，如果岐義詞的意義具有不同的詞性很容易透過詞性標記程式辨識出不同的意義。而像前面的例子 bank 不同的意義如銀行，河堤，庫都是名詞，辨識的困難度增高許多。我們所使用的訓練語料 Senseval-2 English lexical sample，是在 2001 年所發布，語料中包含了 73 個不同的目標詞，詞性有名詞、動詞、形容詞，但同一個目標詞的不同意義詞性都是相同的，對於詞義辨識的演算法形成很大的挑戰。

早期詞義辨識的演算法大都利用利用辭典的定義、或同義詞辭典(thesaurus)的語義分類訊息。例如 Lesk（1986）判斷目標詞的語境與辭典的哪

一個意義的定義最接近，所採用的相似度計算方式以兩者相同的非功能詞的數目為主。Walker (1987)則利用同義詞辭典(thesaurus)當中的語義類別。這些演算法跟目前常用的機器學習演算法相比正確率低許多。

機器學習方法主要可分為監督式(supervised learning)及非監督式(unsupervised learning)。兩者的差別在於前者的訓練語料有標記答案的而後者沒有，我們所採用的方法是監督式的方法。無論是哪一種機器學習的詞義辨識演算法都需要利用語境的訊息。例如 Purandare and Pedersen (2004) 採用非監督式的方法，從沒有標示詞義純文字語料抽出語境並將機讀辭典 Wordnet 裡面不同詞義的定義去除功能詞後建立共現矩陣(co-occurrence matrix)，利用 Singular Value Decomposition (SVD)將維數降到 100，最後用 Latent Semantic Indexing (LSI)找出某一句中的目標詞最有可能的詞義。Jurafsky and Martin (2000)將常用的語境特徵分成兩類。一類是搭配語特徵(collocational features)，另一類是 bag of words information。兩者的最大差別在於後者只考慮某些詞在目標詞左右一定範圍的詞有沒有出現，不考慮這些詞彼此或跟目標詞前後的關係，而前者則納入與目標詞前後相對位置的訊息，甚至用語法剖析器得到語法依存關係。

詞義辨識方法除了可以利用 Semantic Concordancer 或 Senseval 這些有標示詞義的語料之外，還可以利用 pseudoword 或雙語語料。pseudoword 是 Gale et al. (1992)和 Schutze(1992)為了省去標示詞義所需的大量人力與時間所創造出來的方法。透過人造的岐義詞如 banana-door，將語料中所有出現 banana 或 door 都代換成 banana-door，這樣就可以得到類似人工標記詞義的訓練語料。此外，某一個有岐義的詞在另一個語言通常沒有岐義，例如英文的 duty 有兩個意義，但在中文裡則由海關和責任兩個詞來表達。Brown et al. (1991) 及 Gale et al. (1992)利用這個特性，以英法雙語語料庫作為訓練語料，採取目標詞左右若干詞（例如 50 個詞）構成一個語境向量(context vector),再利用 Bayesian

classification 來選擇在某一個語境當中哪一個詞義的機率最大。我們也採用 Bayesian classification 但搭配不同的特徵。Bayesian classification 的概念是目標詞周圍的詞會反映出目標詞的意義，因此將周圍的詞以及目標詞做統計再利用機率選擇詞義，在第三節中會有詳細的介紹。

Yarowsky (1995)注意到在某一篇文章中一個目標詞的詞義通常是固定某一個詞義(One sense per discourse)。且目標詞的搭配語提示了這個目標詞的詞義(One sense per collocation) 。本文所採用搭配語作為機器學習演算的法特徵受到 Yarowsky (1995)的啟發。Lin (1997)有鑑於以機器學習分類器(classifier)來辨識詞義需為不同的詞分別訓練出不同的分類器，頗不方便，因此提出一種使用同一種知識來源(knowledge source)的方法。他利用自己所發展的 MINIPAR 英文剖析器得到的語法依存關係(dependency relations)，如動詞與受詞的關係作為機器學習演算法的特徵。比較特別的地方在於他的方法不需要標示詞義的語料，而是利用相同語意的詞會出現在具有相同的依存關係所組成的局部語境(local context)。Lin (1997) 的正確率達到與其它機器學習演算法相同水準。有關於特徵的選取，Le and Shimazu (2004)針對英文詞義辨識提出數個特徵並以 Forward Sequential Selection Algorithm 來得到最佳的特徵組合。

除了上面介紹的方法，還有許多詞義辨識的方法，例如利用 mutual information 的 Flip-Flop algorithm (Brown et al. (1991)), 使用 decision list (Yarowsky (1994))等，限於篇幅無法一一介紹。近幾年詞義辨識的演算法除了 Naïve Bayes 之外，越來越多人使用 Maximum Entropy，Support Vector Machine，及 Conditional Random Field 等較新的機器學習演算法。