詞義辨識:機器學習演算法特徵的選取與組合

高紹航
denehs@gmail.com
臺灣大學資訊工程學系

高照明
zmgao@ntu.edu.tw
臺灣大學外國語文學系

摘要

詞義辨識(Word Sense Disambiguation, WSD)是自然語言處理中很重要的一環，本文利用 Naïve Bayes 的分類器(classifier)選用目標詞的詞性，相鄰詞及詞性，搭配語，語法依存關係及語義特徵等共九項特徵，以 Le and Shimazu (2004)所提出的 Forward Sequential Selection Algorithm 來得到最佳的特徵組合。我們以 Senseval-2 中的 English lexical sample 做爲訓練語料以及測試語料，得到 61.2%的正確率，與 Senseval-2 參賽第一名的隊伍 64.2%的正確率相差 3%，與第 4 名史丹福大學的參賽隊伍相差 0.5%。

關鍵詞：詞義辨識(word sense disambiguation)，搭配語(collocation)，語法依存關係(dependency relations)，sketch engine，Stanford parser，Hownet，Naïve Bayes，Forward Sequential Selection Algorithm

一、前言

一個英文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。

Senseval-2 的訓練以及測試語料是以 XML 的型式儲存，以下是一筆訓練語料的範例：

```
<instance id="art.40001" docsrc="bnc_ACN_245">
<answer instance="art.40001" senseid="art%1:06:00::"/>
<context>
Their multiscreen projections of slides and film loops have featured in orbital parties, at the
```

Astoria and Heaven, in Rifat Ozbek's 1988/89 fashion shows, and at Energy's recent Docklands all-dayer.
From their residency at the Fridge during the first summer of love, Halo used slide and film projectors to throw up a collage of op-art patterns, film loops of dancers like E-Boy and Wumni, and unique fractals derived from video feedback.
&bquo;We're not aware of creating a visual identify for the house scene, because we're right in there.
We see a dancer at a rave, film him later that week, and project him at the next rave.&equo;
Ben Lewis Halo can be contacted on 071 738 3248.
<head>Art</head>you can dance to from the creative group called Halo
</context>
</instance>

語料中目標字會用<head>以及</head>標出，測試語料格式與訓練與料相似，其差別在於沒有 senseid 的標記。

二、文獻回顧

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爲主。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（請參考表 16 Senseval-2 詞義辨識競賽各個方法的正確率）

機器學習方法主要可分爲監督式(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)爲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爲訓練語料，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爲機器學習演算的法特徵受到 Yarowsky (1995)的啓發。Lin (1997)有鑑於以機器學習分類器(classifier)來辨識詞義需爲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爲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。本文採用語法依存關係作爲詞義辨識的特徵源自於 Lin (1997)的想法。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，本文採用 Le and Shimazu (2004)所提出的 5 個特徵另外加上 4 個特徵，並仿照 Le and Shimazu (2004)所使用的 Forward Sequential Selection Algorithm 得到最佳特徵的組合。

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。本文所選取特徵和組合的方法也可以與這些方法一起使用。

三、我們採取的方法

（一）、Bayesian Classification

在我們的實驗中，我們採用 Bayesian Classification 搭配多種特徵的方法，下面簡述 Bayesian Classification。

假設我們現在要對一個目標詞做詞義辨認，該目標詞的詞義有 k 個，依序是$s_1, s_2, \ldots, s_k$，則目標就是要找出一個s'，使得$P(s'|c)$爲最大，c 是目標詞所含有的某種特徵。根據貝式定理，可以得到如下的等式：

$$P(s_k|c) = \frac{P(c|s_k)}{P(c)}P(s_k)$$

因此

$$s' = \arg\max_{s_k} P(s_k|c)$$

$$= \arg\max_{s_k} \frac{P(c|s_k)}{P(c)}P(s_k)$$

$$= \arg\max_{s_k} P(c|s_k)\,P(s_k)$$

$$= \arg\max_{s_k} [\log P(c|s_k) + \log P(s_k)]$$

我們所有的實驗都是使用這個方法來作詞義辨認，差別是在於選取的特徵的不同。

（二）、Forward Sequential Selection Algorithm

在特徵的選取方面，由於我們嘗試了很多種特徵，假如特徵有 7 種那麼特徵組合的種類就有 127 種，數量非常的可觀，一個一個將所有的組合做實驗非常沒有效率，因此使用 Le and Shimazu (2004)所提出的 Forward Sequential Selection 演算法來挑選特徵。這個方法大致上是先令一個特徵的集合 S 爲空集合，首先挑一個最好的特徵放進 S 中，接著將每一個特徵都放進 S 中看哪個得到的正確率最高來決定第二個要放入 S 中的特徵，如此反覆直到最後正確率不再增加爲止，最後集合 S 中的特徵就會是一個很不錯的特徵組合，雖然未必真的是最佳解但應用在英文詞義辨認的特徵選取上與真正最佳解的差異非常小。

（三）、特徵

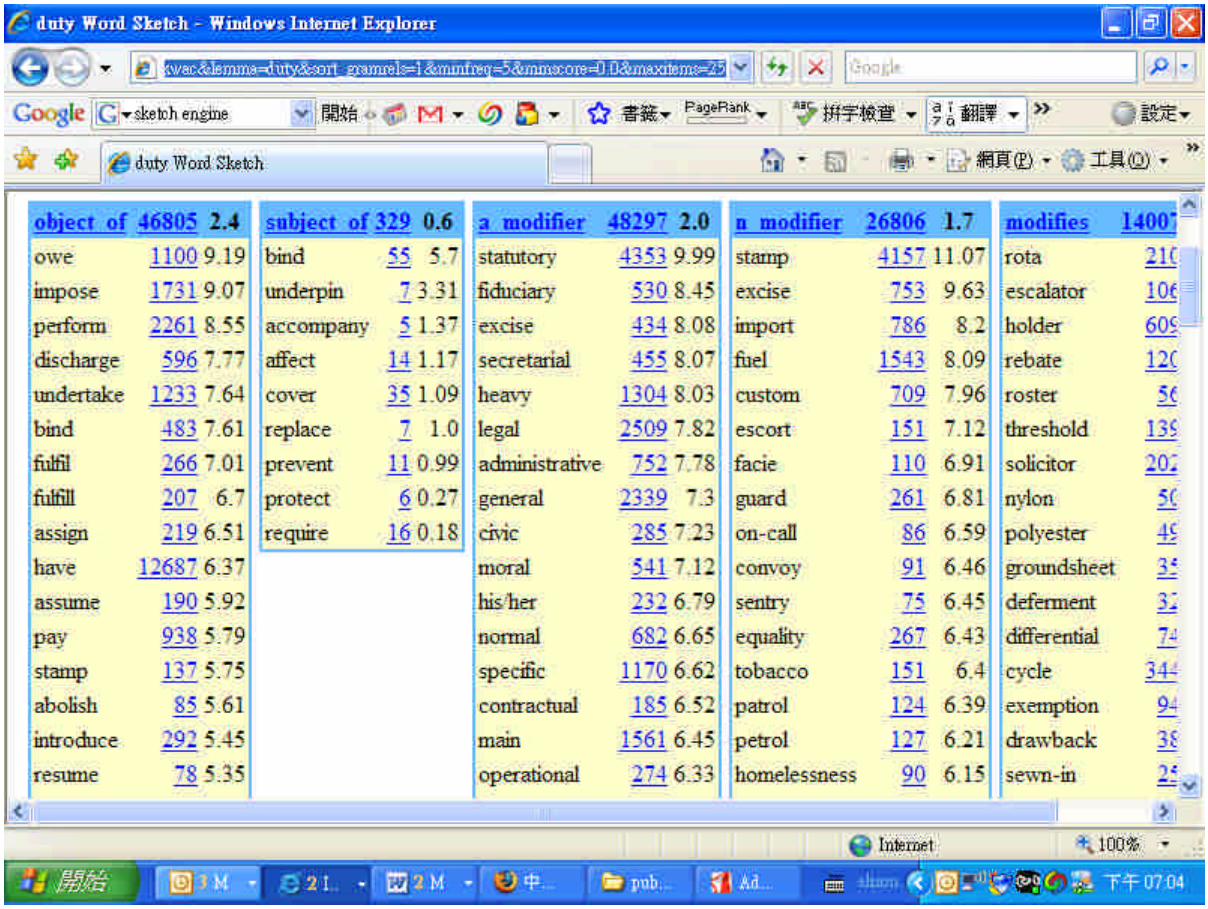
我們一共嘗試了 9 種特徵，分別以 F1 到 F9 命名之，前五個主要是針對目標詞周圍的詞以及其詞性，這 5 個特徵都是 Le and Shimazu (2004)所使用的特徵，第六個以及第七個則著重在詞的依存關係，例如主詞與動詞，動詞與受詞的關係等等，而最後兩個則是利用 HowNet 的取目標詞的前後兩個詞以及與目標詞有依存關係的 HowNet 義元(語義特徵)當作特徵，在此一一介紹。

F1 是直接把目標詞周圍的詞做爲特徵，但是會排除一些如 is, a 之類的功能詞(stop words)。
F2 也是目標詞周圍的詞，但會加上位置的資訊，例如目標詞是 art 時，〝The art of design〞中會被取出的特徵會是{(The, -1), (of, 1), (design, 2)}。
F3 跟 F2 類似，但不同的是 F3 是取出詞性。
F4 則是目標詞與周圍詞的組合，同樣以〝The art of design〞爲例，會被取出的特徵有{The-art, art-of, The-art-of, art-of-design, The-art-of-design}。
F5 與 F4 類似但取詞性組合。

F6 則是利用 Sketch Engine，將可能與目標詞有語法搭配關係的詞列爲特徵。
F7 是利用 Stanford Parser，將 Stanford Parser 所剖析出的與目標詞有依存關係的詞以及依存關係的類別列爲特徵。
F8 是取目標詞前後兩個詞的 HowNet 義元做爲特徵。
F9 是先利用 Stanford Parser 找出與目標詞有依存關係的詞，再取出其 HowNet 義元做爲特徵。

我們使用 Sketch Engine (Kilgarriff et al. (2004))找出跟目標詞具有語法搭配關係的所有詞。圖一是利用 Sketch Engine (http://www.sketchengine.co.uk/)的 word sketch 查詢 duty 個目標詞的輸出，object_of 這一欄表示目標詞可以作爲這些詞的受詞的搭配語，subject_of 表示目標詞可以作爲這些詞的主詞的搭配語, a_modifier 是可以修飾這個目標詞的形容詞，n_modifier 是可以修飾這個目標詞的名詞, modifies 則是可以被這個目標詞修飾的詞。我們選擇的英文語料超過 20 億，如此龐大的語料可以確保得到大部分的搭配語。

圖一 Sketch Engine Word Sketch 的輸出結果



Stanford Parser 是史丹福大學 Klein and Manning (2003)發展出來的多國語言剖析器，只要輸入符合 Pen Treebank 格式的語法樹庫，即可自動從語法樹庫中訓練得到該語言的語法剖析器。下面的例子是 Stanford Parser 的輸出結果。除了標示詞性，語法結構，

最特別的是還將語法的依存關係列出來,例如,nsubj 表示動詞和主詞的關係，dobj 表示動詞和受詞的關係，advmod 表示動詞和副詞修飾語的關係，amod 表示名詞和名詞修飾語的關係。必須強調的是 Stanford Parser 的語法依存關係是從語法樹庫歸納出來後利用 regular expression 抽取出來的，因此即使語法剖析的結果正確，語法依存關係不一定正確。下面是 Stanford Parser 的輸出結果。

The/DT government/NN first/RB established/VBD modern/JJ criminal/JJ investigation/NN system/NN in/IN 1946/CD ./.

```
(ROOT
  (S
    (NP (DT The) (NN government))
    (ADVP (RB first))
    (VP (VBD established)
      (NP
        (NP (JJ modern) (JJ criminal) (NN investigation) (NN system))
        (PP (IN in)
          (NP (CD 1946)))))
    (. .)))
```

```
det(government-2, The-1)
nsubj(established-4, government-2)
advmod(established-4, first-3)
amod(system-8, modern-5)
amod(system-8, criminal-6)
nn(system-8, investigation-7)
dobj(established-4, system-8)
prep(system-8, in-9)
pobj(in-9, 1946-10)
```

知網 Hownet(http://www.keenage.com)是由董振東所發展出來（參考 Dong and Dong (2006)）。Hownet 架構不同於 Wordnet，Wordnet 基本上是一個詞彙網路，同樣語意的詞屬於同一組的 synset，裡面的定義，例句都相同。Wordnet 裡面包含的詞彙語意關係包括上位詞，下位詞等。Hownet 則利用抽象的義元作爲表達所有概念的工具和單位。Hownet 包含的訊息相當的多，是一個中英雙語的知識庫，包括義元，語意角色，上下位關係，部件與整體關係等等語意訊息。義元類似一個語意特徵。Hownet 對於醫生(doctor)的義元表示法爲

{human| 人 :HostOf={Occupation| 職 位 },domain={medical| 醫 },{doctor| 醫

治:agent={~}}}

Hownet 裡面的訊息表示醫生是一個人，具有職位，屬於醫學領域，醫生在醫治這個事件裡扮演主事者的語義角色。在我們的實驗中，我們只使用 Hownet 表示法當中第一個義元，例如：doctor 的第一個義元是 human。對於名詞而言，第一個義元相當於這個詞的語意類別或本體 ontology。

四、實驗結果

在 F1~F6 中，都必須取一個 Window Size，否則會導致特徵和目標詞以及詞義的相關聯性表現不出來，因此這六種特徵都會有 Window Size 的實驗。而我們的實驗是對各種特徵先獨立的來做詞義辨認以得到各種特徵的最佳參數，每種特徵都最佳化以後，最後再使用 Forward Sequential Selection Algorithm 來決定要採用哪些特徵。處理語料以及辨認程式是以 Perl 及 C++寫成。

（一）、F1

F1 是最簡單的直接把目標詞周圍的詞做爲特徵，但是會排除一些如 is, a 之類的 stop words，原因是 stop words 通常對於辨認一個詞的詞義沒有什麼幫助。表一顯示 F1 的最佳 Window Size 爲 3。

表一、F1 Window Size 實驗結果

Window Size	正確率(%)
1	52.7
2	54.2
3	54.6
4	54.6
5	54.1

（二）、F2

F2 也是目標詞周圍的詞，但會加上位置的資訊，會記錄某個詞是出現在目標詞的什麼位置，表二顯示 F2 的最佳 Window Size 爲 1

表二、F2 Window Size 實驗結果

Window Size	正確率(%)
1	54.9
2	53.6
3	51.1
4	47.9

（三）、F3

F2 是目標詞周圍的詞，但會加上位置的資訊， F3 跟 F2 類似，但不同的是 F3 是取出詞性。

表三、F3 Window Size 實驗結果

Window Size	正確率(%)
1	44.6
2	35.5
3	30.7
4	27.7

因此 F3 的最佳 Window Size 爲 1

（四）、F4

F4 則是目標詞與周圍詞的組合，同樣以〝The art of design〞爲例，會被取出的特徵有{The-art, art-of, The-art-of, art-of-design, The-art-of-design}。

表四、F4 Window Size 實驗結果

Window Size	正確率(%)
1	48.2
2	56.9
3	57.8
4	57.8
5	57.8

因此 F4 的最佳 Window Size 爲 3。

（五）、F5

F5 則是目標詞與周圍詞性的組合。

表五、F5 Window Size 實驗結果

Window Size	正確率(%)
1	48.2
2	52.1
3	53.8
4	54.2
5	54.1

因此 F5 的最佳 Window Size 爲 4。

（六）、F6

F6 則是透過 Sketch Engine 將可能與目標詞有依存關係的詞列爲特徵。Sketch Engine 在使用時有數個參數可以做調整，分別是所要包含的依存關係種類、minimum salience，在做 Window Size 實驗時，minimum salience 設爲 0、依存關係種類爲全部，而在做 minimum salience 時 Window Size 設爲 5、依存關係種類爲全部，在做依存關係種類選擇時，minimum salience 設爲 0、window size 設爲 5。

表六、F6 Window Size 實驗結果

Window Size	正確率(%)	Window Size	正確率(%)
1	50.5	11	51.6
2	51.1	12	51.4
3	51.6	13	51.4
4	51.8	14	51.1
5	**52.0**	15	50.8
6	**52.0**		
7	51.8		
8	51.5		
9	51.4		
10	51.5		

因此 F6 的最佳 Window Size 爲 5。

表七、F6 minimum salience 實驗結果

Minimum salience	正確率(%)
0.0	52.0
1.0	51.8
2.0	51.8
3.0	51.3

因此 F6 的最佳 Minimum Salience 約爲 0.0。

對於依存關係的選擇，則是使用 Forward Sequential Selection Algorithm 來選出最好的組合。

表八、F6 依存關係組合選擇(第一步)

Type	正確率(%)	Type	正確率(%)
Object	49.2	and/or	49.4
object_of	47.5	pp*	49.4
Subject	48.4	possessor	46.6
subject_of	48.0	possessed	47.6
a_modifier	48.1	Modifier	48.4
n_modifier	49.0	part*	48.5
Modifies	**50.1**	*comp_of	48.3
		*comp	48.2

在這步中選擇了 modifies

表九、F6 依存關係組合選擇(第二步)

Type	正確率(%)	Type	正確率(%)
Object	**51.1**	and/or	50.8

object_of	50.1	pp*	50.9
Subject	50.2	possessor	50.1
subject_of	50.1	possessed	50.0
a_modifier	50.3	Modifier	50.2
n_modifier	50.7	part*	50.2
*comp	50.1	*comp_of	50.1

在這步中選擇了 object

表十、F6 依存關係組合選擇(第三步)

Type	正確率(%)	Type	正確率(%)
object_of	51.2	and/or	51.5
Subject	51.1	pp*	51.4
subject_of	51.1	possessor	51.1
a_modifier	51.3	possessed	51.0
n_modifier	**51.5**	Modifier	51.2
Comp	51.1	Part	51.1
		comp_of	51.2

在這步中選擇了 n_modifier

表十一、F6 依存關係組合選擇(第四步)

Type	正確率(%)	Type	正確率(%)
object_of	51.6	and/or	51.7
Subject	51.5	pp*	51.6
subject_of	51.5	possessor	51.5
a_modifier	**51.7**	possessed	51.5
comp_of	51.4	Modifier	51.4
Comp	51.4	Part	51.3

在這步中選擇了 a_modifier

表十二、F6 依存關係組合選擇(第五步)

Type	正確率(%)	Type	正確率(%)
object_of	51.6	and/or	**51.9**
Subject	51.7	pp*	51.8
subject_of	51.7	possessor	51.6
comp_of	51.7	possessed	51.7
Comp	51.8	Modifier	51.8
		Part	51.7

在這步中選擇了 and/or

表十三、F6 依存關係組合選擇(第六步)

Type	正確率(%)	Type	正確率(%)

object_of	51.9	pp*	51.9
Subject	51.9	possessor	51.9
subject_of	51.9	possessed	51.9
comp_of	51.9	Modifier	**52.0**
Comp	51.9	Part	51.9

在這步中選擇了 modifier

表十四、F6 依存關係組合選擇(第七步)

Type	正確率(%)	Type	正確率(%)
object_of	52.0	pp*	51.9
Subject	52.0	possessor	51.9
subject_of	51.9	possessed	51.9
		Part	52.0
		comp_of	52.0
		Comp	52.0

在這步中可以看到無論加進哪種依存關係正確率都不再上升了，因此最後所找到的最佳依存關係組合爲{ modifies, object, n_modifier, a_modifier, and/or, modifier}。這個結果顯示主詞的特徵對於詞義辨識而言不是很重要。最重要的是修飾語和受詞。

（七）、F7

F7 是利用 Stanford Parser 所剖析出與目標詞有依存關係的詞以及依存關係的類別（例如：object_of, modifies）列爲特徵。準確率是 54.6%

（八）、F8

F8 是取目標詞前後兩個詞的 HowNet 義元做爲特徵。準確率是 47.2%

（九）、F9

F9 是先利用 Stanford Parser 找出與目標詞有依存關係的詞，再取出其 HowNet 義元做爲特徵。準確率是 54.1%

（十）、特徵選取

採用 Forward Sequential Selection Algorithm 來做特徵選取。

表十五、特徵選取結果

Step	F1	F2	F3	F4	F5	F6	F7	F8	F9
1st	54.6	54.9	44.6	**57.8**	54.2	52.0	54.6	47.2	54.1
2nd	58.9	58.5	56.2		56.8	58.2	**59.6**	56.8	59.2
3rd	**60.7**	60.1	58.2		58.1	60.2		59.1	59.7
4th		**61.2**	60.1		58.8	60.6		60.4	60.7

5th			60.3		59.4	60.8		60.4	61.1

每個 step 會有一欄是粗體, 代表該 step 選取的特徵。例如，第一步選出 F4，接下選出 F7,換言之，第二步的 F1 其實是代表 F4+F1, F2 代表 F4+F2..依此類推。而第二步驟結束後選取的特徵就是 F4+F7。在第三步驟時的 F1 是代表 F4+F7+F1，以此類推。因此最好的特徵組合爲目標詞週圍的三個詞、目標詞週圍一個詞及其位置關係、目標詞與周圍三個詞的連續組合、以及利用 Stanford Parser 所得到與目標詞有依存關係的詞，正確率是 61.2%。

由實驗結果可看出，由於 senseval-2 的岐義目標詞詞性都一樣，採用詞性相關的特徵對於詞義辨認沒有什麼幫助，較有幫助的是目標詞周圍的詞以及與其有依存關係的詞。

五、結論

下表是 Senseval-2 當時的結果，這份結果所使用的訓練及測試語料和我們使用的是相同的：

表十六、Senseval-2 English Lexical Sample Result

準確度	系統	準確度	系統
64.2	JHU (R)	51.2	Baseline Lesk Corpus
63.8	SMUls	50.8	Duluth B
62.9	KUNLP	49.8	UNED - LS-T
61.7	Stanford - CS224N	47.6	Baseline Commonest
61.3	Sinequa-LIA - SCT	43.7	Baseline Grouping Lesk Corpus
59.4	TALP	42.7	Baseline Grouping Commonest
57.1	Duluth 3	41.1	Alicante
56.8	JHU	26.8	Baseline Grouping Lesk
56.8	UMD - SST	24.9	IRST
56.4	BCU - ehu-dlist-all	23.3	BCU - ehu-dlist-best
55.4	Duluth 5	23.0	Baseline Grouping Lesk Def
55.0	Duluth C	22.6	Baseline Lesk
54.2	Duluth 4	18.3	Baseline Grouping Random
53.9	Duluth 2	16.3	Baseline Lesk Def
53.4	Duluth 1	14.1	Baseline Random
52.3	Duluth A		

由表中數據可看出，如果對每個目標詞隨機選一個意義的話準確度是 14.1%、選最常見的意義的話是 47.6%。而我們的結果 61.2%遠大於 baseline 的數據而且也比大部分的系統好，表示這些特徵應用在詞義辨認中是有效的。我們實驗的結果顯示主詞的特徵對於詞義辨識而言不是很重要。最重要的是修飾語和受詞。雖然與 Senseval 2 參賽裡面最好的系統正確率還相差 3%,我們正在實驗其它重要的特徵（如 Wordnet 的 synset, lexicographical file, 及定義）並嘗試用其它機器學習演法如向量支撐機(SVM)或 CRF 希

望進一步提升正確率。

致謝

本研究得到下列國科會計畫經費補助，特此致謝。「詞彙語意關係之自動標注─以中英平行語料庫爲基礎(3/3)」 NSC 93-2411-H-002-013「中英平行句法樹庫的建立與英漢結構對應演算法的研究(I)(II)」NSC94-2411-H-002-043 NSC95-2411-H-002-045-MY2

參考文獻

Brown, Peter et al. (1991) Word sense disambiguation using statistical methods. In ACL 29, pp. 264-270.

Dong, Zhendong and Dong, Qiang. (2006) Hownet and the Computation of Meaning. World Scientific.

Gale, William, Church, Kenneth, and Yarowsky, David. (1992) A method of disambiguating word senses in a large corpus. Computers and the Humanties 26:415-439.

Jurafsky, Daniel, and James H. Martin. (2000) Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall.

Klein, Dan. and Manning, Christopher. (2003) Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

Le, Cuong Anh and Shimazu, Akira. (2004) High WSD Accuracy Using Naïve Bayesian Classifier with Rich Features. PACLIC 18, Tokyo. http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/564/1/oral-8.pdf

Lesk, Michael. (1986) Automatic Sense Disambiguation: How to tell a pine cone from an ice cream cone. In Proceedings of the 1986 SIGDOC Conference, pp. 24-26, New York. Association for Computing Machinery.

Lin, Dekang . (1997). Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity In *Proceedings of ACL-97,* Madrid, Spain. July, 1997.

Manning, Christopher, and Schutze, Hinrich. (1999) Foundations of Statistical Natural Language Processing. MIT Press.

Patwardhan, Banerjee, and Pedersen (2005) SenseRelate::TargetWord - A Generalized Framework for Word Sense Disambiguation. Appears in the Proceedings of the Twentieth National Conference on Artificial Intelligence, July 12, 2005, Pittsburgh, PA. (Intelligent Systems Demonstration)

Purandare and Pedersen (2004) Improving Word Sense Discrimination with Gloss Augmented Feature Vectors. Appears in the Proceedings of the Workshop on Lexical

Resources for the Web and Word Sense Disambiguation, November 22, 2004, Puebla Mexico.

Yarowsky, D. (1994) Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French." In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics.* Las Cruces, NM, pp. 88-95.

Hownet http://www.keenage.com/

senseval-2 http://193.133.140.102/senseval2/

Sketch Engine http://www.sketchengine.co.uk/

Stanford Parser http://www-nlp.stanford.edu/downloads/lex-parser.shtml