基於非監督式詞義消歧之日語旅遊意見詞翻譯

Japanese Opinion Word Translation Based on Unsupervised Word Sense Disambiguation in the Travel Domain

黃俊瑋 [1] 高嘉珮 [2] 陳駿瑜 [3] 蔡宗翰 [4]
Jyun-Wei Huang, Kao Chia Pei, Chun-Yu Chen, Richard Tzong-Han Tsai
元智大學資訊工程學系
Department of Computer Science and Engineering
Yuan Ze University
{s976017, s951559}@mail.yzu.edu.tw[1,3]
{chia, thtsai}@saturn.yzu.edu.tw[2,4]

摘要

本論文提出了一個加上特徵詞字典與相依性關係詞的非監督式意見詞翻譯方法，我們的方法包含了在測試階段前的語料庫準備、特徵詞字典產生以及加權方式，不同於當今機器翻譯的方式，不須另外準備平行語料庫、標記語料庫甚至是剖析樹語料庫。本研究僅需一個可信賴的線上雙語字典即可完成，所需成本極小但是效果較現時諸多線上翻譯系統來得準確。此外也可以從評論中擷取出關鍵資訊，幫助使用者更加了解評論之內容。我們設計三種組態，並挑選四個在日語中出現頻率非常高的意見詞進行實驗。由實驗結果顯示，相依性關係與特徵詞字典確實對於意見詞翻譯是有效的，而此種方法除了能解決一字多義現象也能提高準確率，幫助使用者了解日語之意見內容助其進行決策。

Abstract

This paper proposes a Japanese opinion word translation method based on unsupervised word sense disambiguation. The method comprises the corpus preparation, opinion word dictionary construction, and weighting method. Different from the machine translation, our method does not need parallel corpora, tagged corpora or parsing tree banks. Our method is low-cost but effective, and requires a well-made bilingual dictionary only. Besides, our method can extract key information from the opinions to help users understand the opinions. We construct four configurations and evaluate our method on four Japanese opinion words with high frequency. The evaluation result shows that the dependency grammar and opinion word dictionary is effective on opinion word translation. Our method can deal with the translation disambiguation problem and improve the translation precision to help user realize Japanese opinions.

關鍵詞：詞義消歧，意見詞，字詞翻譯，非監督式

Keywords: Word Sense Disambiguation, Opinion Word, Word Translation, Unsupervised

一、緒論

隨著網路的蓬勃發展，越來越多的使用者會在網路上發表對產品的評論，或者對於美食

Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 157-171, Puli, Nantou,Taiwan, September 2010.

的食記，對於旅遊的遊記，對於住宿旅館的評價，除了分享與發表之外，也透過閱讀評論，來作為是否把該目的物加入考量的一個方式。像是目前在多國規模領先的搜尋引擎 Google 也漸漸重視到這塊，在搜尋選項中可以選擇「討論」這個項目。

且受限於語言的隔閡，多數消費者無法理解以外語撰寫的評論。除了較普遍的第二外語英文之外，大多數的華人無法了解檢索出來的外語網頁。以旅遊住宿資訊為例，台灣有背包客網站（http://www.backpackers.com.tw/），但是如果今天有一家日本新開的旅館，而在台灣的背包客網站也許要過好幾個月後才會有人分享，甚至若一直沒有台灣人去住過，也許永遠看不到那家旅館的評論資訊。另外，對於小眾的旅行者而言，如：走訪當地的鄉土民情、登山、釣魚等，當地語言的相關資訊或者評論也較完整。

而在意見探勘（Opinion Mining）的研究領域上，大致可以分成幾項任務來處理[1]，分別是：

- 從整篇評論的觀點：
  - 判斷整篇評論的情緒傾向
- 從句子的觀點：
  - 判斷句有主觀性/意見性的句子
  - 判斷句子的情緒傾向
- 從特徵的觀點：
  - 判斷由什麼提出的評論（例如：評論者）
  - 判斷意見詞的情緒傾向
  - 聚集分類特徵

其中在從特徵的觀點上，還有判斷意見詞以及判斷特徵詞的問題存在。而本研究挑選意見詞來判斷詞義主要基於在評論句當中最重要的資訊就是特徵詞以及意見詞兩者，當使用者知道句子中的特徵詞與意見詞即可以做決策是否要購買或者使用等等。且在意見詞方面，大多數是形容詞居多，形容詞在經過線上字典查詢過後，翻譯結果的差異性比較大。再加上意見詞用詞程度也會影響使用者的決策，因為每個人的選擇不同，可以接受好壞程度的範圍不同，有些人也許只去評論最好的地方，有些人則是覺得普通就可以接受了。因此，在翻譯評論句中的意見詞詞義上更顯得重要，也是最容易造成使用者誤解的部分。

本論文有別於現今在自然語言處理的問題之機器翻譯技術，機器翻譯技術可初略的分為統計式（Statistical Machine Translation）與規則式（Rule Base）兩類，早期的研究者較注重規則式的翻譯方法，但是由於網際網路上的語料越來越多，目前統計式的方法比較受到重視。

以往，採用規則比對方式是採用人為給定規則權重的方式，以進行剖析或詞性標記等動作。由於權重是人為給定的，這使得每條規則的影響力落入主觀的判斷之中，若在語料庫內沒有包含此規則的現象，也很難去完全的包含取到所有的規則。而統計式的翻譯所依靠的主要是語料庫，而非語法規則，機器翻譯系統的語料庫通常稱為平行語料庫（Parallel Corpus），甚至，有些語料庫還記載著詞性標記的資訊的標記語料庫（Tagged Corpus），或記載著剖析樹的結構資訊為樹庫（Tree Bank）。然而，建立與取得這些語料庫都相當花費人工與時間，準確率也並不一定較佳。

因此，就在意見分析上的重點與特色以及機器翻譯系統所需的建構成本較大。本論文著重在非監督式的方式，利用一個線上辭典以及網路上的語料，再經由共現的情況與其他

特徵資訊來判斷評論句中最關鍵的意見詞的詞義。

二、文獻探討

在過去的十五年，自然語言處理領域中的一個顯著改變，即為從人工製作的系統轉變到自動分類的方法[2]。對於機器學習技術的興趣有如此巨大的改變，可以從監督式方法應用到詞義消歧（Word Sense Disambiguation）的問題可見一斑。監督式的詞義消歧方法利用機器學習的技術與人工所標註的資料製成分類器。通常來說，此分類器（亦稱為文字專家，Word Expert）主要著重於單一字，並扮演著分類的角色來對每種情況指定適當的詞義。而用來學習分類器的訓練集（Training Set）通常是由許多人工標註後的例子以及目標詞的詞義等資訊所組合而成。根據 McCarthy[3]學者的研究，監督式方法在訓練資料充足的情況下，正確率會高於其他的詞義辨識技術。

例如在上下文特徵方面，常用的特徵有 Co-occurrence Word Frequency、Part-of-Speech、Syntactic、Collocation 等，針對每一個特徵分別設計系統，再將所有系統以權重值結合，其執行效能比起一次擷取多項特徵並設計一套複雜系統的方式要來的好[4-5]。另一方面，透過交錯驗證（Cross Validation）的方式找到最佳的合併特徵，再將合併特徵丟入貝氏分類器辨識目標詞之詞義，在辨識效果上有不錯的系統表現[6]。監督式學習法雖然需仰賴人工標示語料庫作為學習之依據，但執行效能與正確性都比其他方式來得優越，因此廣泛為大部分研究所使用。

非監督式學習法是直接由原始語料庫找尋目標詞彙所含的上下文特徵，直接分類目標詞彙之詞義的一種詞義辨識技術。這樣的方式，並不需要任何事先標示詞義的資料集，取而代之的是以較複雜的數學模組做相似度計算，分類目標詞彙。也因此，學者 Gale[7] 也提到非監督式方法擁有克服知識獲取瓶頸的潛能。

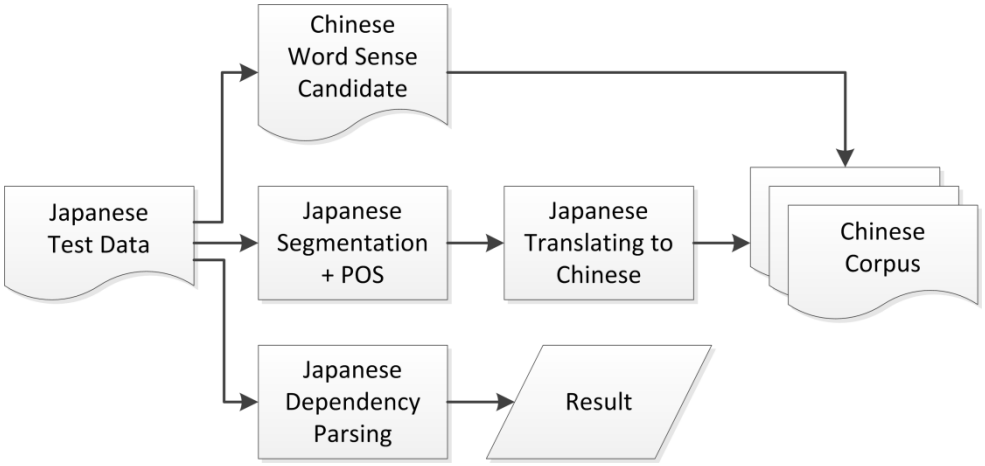
McCarthy 等作者的研究中，先由原始語料庫對詞彙分類以推導同義詞典，再由同類詞彙群選出與目標詞彙最相似的 k 個詞彙，並由 WordNet 擷取目標詞彙的每個詞義與 k 個相似詞彙的最相似詞義做相似度成績加總，以成績最高分之詞義作為目標詞彙之詞義[8]。在中文詞義辨識處理上，Lu 等學者以同義詞詞林[9]收集與目標詞同一詞義之詞彙作為虛擬詞彙集（Pseudo Word），再由原始語料庫中，針對同一詞義統計虛擬詞彙集的出現機率值並加上周邊詞彙與虛擬詞彙集的共現機率值，再以機率模式之計算標示目標詞彙之詞義[10]。

三、方法

本論文之研究主要著重在僅需準備一個可依賴的線上雙語字典，且不包含任何排名或者出現頻率等排序，即可判斷詞義的趨近非監督式演算法，因此在系統整體流程中並沒有訓練階段。本章節分為三大部分，第一小節為迭代式計算，第二小節與第三小節則為在迭代式計算中所需用到的加權值方式與資料。

（一）、迭代式計算

在本小節主要介紹在迭代式計算中的前處理方式與計算過程，



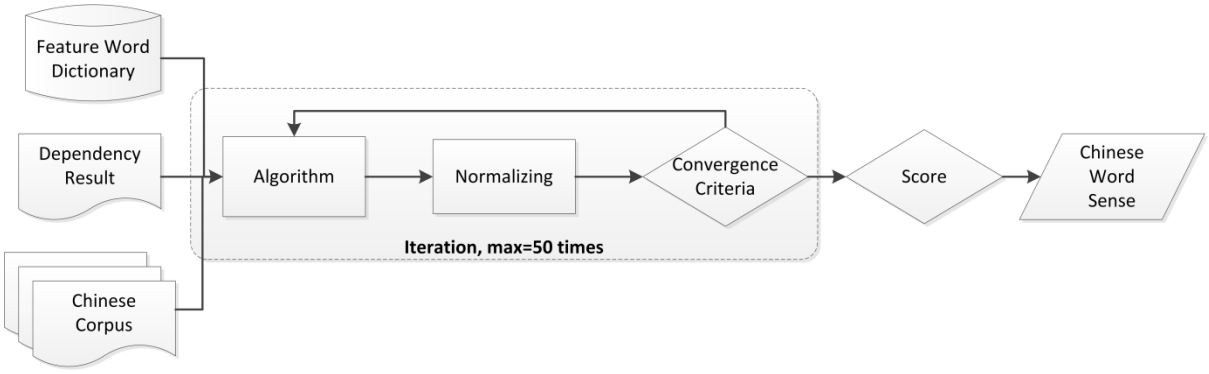
圖一、測試資料前處理

前處理部分如圖一。首先，將需要測試的評論句子取得我們所要辨別的意見詞以及該意見詞的候選詞義，經由日文的斷詞工具 MeCab（http://mecab.sourceforge.net/）進行斷詞以及詞性標註，接著去除我們所定義的一些停用字（助詞、記號、助動詞、接頭詞）。然後把將保留的日文字詞，除了需判斷的意見詞外，都經過 Excite 日中線上字典（http://www.excite.co.jp/dictionary/japanese_chinese/）的翻譯，取得周邊相關字的中文翻譯詞表。再將周邊相關字的中文翻譯與意見詞的候選翻譯詞去線上搜尋引擎 Google 抓取網頁片段摘要（Snippet）回來當作語料庫，抓取時間約在 2010 年初，每筆查詢所抓取的數量為五百筆片段摘要。此階段也先將原測試的評論句子進行在本章第三小節所描述的產生相依性過程步驟，以便於在計算階段能直接取用結果。

而在日中線上翻譯字典當中，我們之所以採用 Excite 辭典，是由於該線上辭典在廣為大多數日本人所使用，並且能提供本研究中所需的中文翻譯詞。而 Yahoo! JAPAN 以及 Google Japan 並沒有提供中日／日中辭典功能，僅有中日／日中翻譯，且它們著重的是在整句翻譯的部分，而不像一般辭典會有多個候選詞出現，所以造成在處理的時候缺少考量未出現的字詞組合，正確性也比較低。

舉例來說，有一日文評論句「写真じゃすごい綺麗な外観だけど。」（照片上看起來外觀非常漂亮。），其中「綺麗」這個日文詞為我們想要判斷詞義的意見詞。首先經由斷詞與詞性標註後成為「写真（名詞）/じゃ（助詞）/すごい（形容詞）/綺麗（名詞）/な（助動詞）/外観（名詞）/だ（助動詞）/けど（助詞）」，然後將一些停用字去除變成「写真（名詞）/すごい（形容詞）/綺麗（名詞）/外観（名詞）」，在將除了「綺麗」之外的其他日文詞進行線上翻譯後取得「照片、相片、可怕、非常、厲害、了不起、外表、外觀」等周邊相關中文翻譯詞。然後再進行製作語料庫的步驟，即是將「綺麗」的翻譯候選詞「漂亮、乾淨」等與周邊相關中文翻譯詞排列組合後去 Google 搜尋引擎抓取前五百筆網頁摘要片段資訊回來，如利用：「（漂亮 AND 照片）」、「（乾淨 AND 照片）」等。

接下來則進入計算分數的過程，如圖二。利用上述前處理過程產生的語料庫、相依性關聯資訊、以及下一小節會介紹到的特徵詞字典，經由公式的計算過後分別會產生每個候選詞義（漂亮、乾淨）的分數，再利用正規化將總和加總改變為 1 或-1，最後判斷是否達到收斂的條件，若沒有的話則繼續計算下去，根據開發資料集（Development set）所計算大多數皆可在 50 次計算內達到收斂的程度。若已經達標準則將高分的作為系統判斷的詞義。



圖二、系統判斷流程

以下將會說明計算意見詞候選翻譯的權重方式，這部分為參考自 C. Monz 與 B. J. Dorr[11]的研究。

首先，第一步為初始步驟。一開始對於每個意見詞或者候選翻譯周邊相關字的翻譯機率是公平的，因此會先給定初始機率，假若意見詞 s_i 有 n 個翻譯，則候選詞義的初始權重 w_T 為 $1/n$。例如「綺麗」若有「漂亮、美麗、乾淨、清潔」四個候選詞義，則每個候選詞義的初始權重皆為 0.25。如公式 3-1 所示：（t 表 s_i 的翻譯候選詞義）

初始步驟（Initialization Step）：

$$w_T^0(t|s_i) = \frac{1}{|tr(s_i)|}, \quad \mathrm{t} \in \mathrm{tr}(s_i) \qquad (0\text{-}1)$$

而在每個候選詞義的初始值給定後，每個候選詞義的權重會基於兩部分的輸入值重新計算：一為周邊關聯字的初始權重$w_T(t'|s_i)$，計算方法同候選詞義初始權重的方式；另一為該候選詞義與周邊關聯字的連結強度$w_L(t,t')$。公式如 3-2 所定義：

迭代式步驟（Iteration Step）：

$$w_T^n(t|s_i) = w_T^{n-1}(t|s_i) + \sum_{t\prime \in inlink(t)} X \cdot w_L(t,t') \cdot w_T(t'|s_i) \qquad (0\text{-}2)$$

其中 *inlink(t)*代表周邊關聯字的翻譯集合，亦即該測試句中若周邊的關聯字有 a_j 與 a_k 兩字，而 a_j 與 a_k 的翻譯為$\{a_{j1},a_{j2},a_{j3},a_{k1},a_{k2}\}$，則這五個字將會與候選翻譯詞 t 進行初始權重計算與關聯強度計算。而其中 X 為本論文貢獻之一，是由特徵詞字典與相依性關係所判斷出來的詞的加權分數，會根據$w_L(t,t') \cdot w_T(t'|s_i)$ 的正負情況給予大於 1 或者介於 1 到 0 之間的加權倍數，若$w_L(t,t') \cdot w_T(t'|s_i)$大於 0，則 X 因子則會大於 1，經由此 X 因子的加權可以把在日文評論使用上與意見詞相關的字詞權重加倍；若$w_L(t,t') \cdot w_T(t'|s_i)$小於 0，經由權重的小部分修正後，可縮小對於相關字詞對於意見詞的負面影響，因為有些是中文網頁在網頁摘要的部分雜訊較多，而造成計算連結強度的部分負面的影響過大。亦即此 X 因子可以使得周邊關聯字的翻譯集合依對於測試句中的意見詞重要性來加權。加權公式如 3-3 所示，其中 *distance* 為相依性關係部分的參數之一，而 F 為特徵詞字典加權變數之一：

X 因子：

$$\mathrm{X}=\begin{cases}1-\left[F+\log\left(1+\frac{1}{distance}\right)\right], & w_L(t,t')\cdot w_T(t'|s_i)<0\\ 1+\left[F+\log\left(1+\frac{1}{distance}\right)\right], & w_L(t,t')\cdot w_T(t'|s_i)\geq 0\end{cases}, \begin{matrix}F=\{0,0.08\}\\ distance\in N\end{matrix} \quad (0\text{-}3)$$

而於公式 3-2 所計算的關聯強度$w_L(t,t')$是在計算選詞義與周邊相關詞中文翻譯的關聯強度，如「漂亮 AND 照片」、「乾淨 AND 照片」等。在本論文利用 Mutual Information[12]方式來計算，在公式 3-4 中，*p(t,t')*是詞 *t* 與詞 *t'*在限定範圍內共同出現的情況，本研究中的限定範圍則以句為單位。

關聯強度：

$$W_L(t,t')=\log_2\frac{p(t,t')}{p(t)\cdot p(t')} \quad (0\text{-}4)$$

再經上述步驟計算每個候選翻譯詞的權重後，把該回合算出來的每個候選翻譯詞的權重分數平均，讓候選翻譯詞的權重分數加起來等於 1 或-1，如公式 3-5。於 C. Monz 與 B. J. Dorr[11]的研究中在正規化計算時並未在分母加上處理絕對值的情況，但在本論文研究範圍內是否加上絕對值處理的效果差異頗大，這部分可歸因於所使用的語料庫不同的關係，C. Monz 與 B. J. Dorr[11]的研究中所使用的語料庫是標準的雙語資料 CLEF 2003，因此語料庫的品質與本研究所使用的網頁摘要語料庫比較起來品質較好些，因此在前步驟計算關聯強度時的分數亦較為正常，較不容易出現負相關情況，本研究則反之。

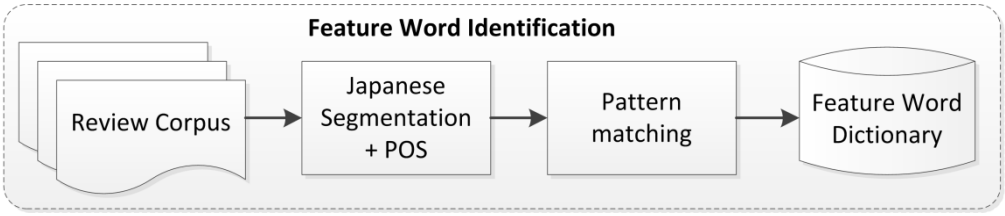
正規化步驟（Normalization Step）：

$$w_T^n(t|s_i)=\frac{w_L^n(t|s_i)}{\left|\sum_{m=1}^{|tr(s_i)|}w_L^n(t_{i,m}|s_i)\right|} \quad (0\text{-}5)$$

經由以上的計算後所取得每回合結束的候選詞義權重分數，再與上一回合的候選詞義相比較是否已達到收斂情況，若無則繼續執行迭代式計算。若已收斂，則將所有候選詞義中最高分詞義所屬的詞義類別判斷為答案。

（二）、特徵詞辨識

在本節介紹特徵詞辨識的方法，圖三為特徵詞辨識流程。本章節所提到的特徵詞即是在意見分析領域中的意見目標（Opinion Target）一詞。首先利用我們所抓取的樂天網站旅遊評論資料，經過斷句、日文斷詞、詞性標註以及停用字（Stop Word）去除的動作，再去做樣式比對（Pattern Matching），而樣式的規則主要經由人工觀察與參考 Kobayashi 的研究[13]整理而得，最後進行過濾的動作後即產生本研究中所需要的特徵詞字典。數量上，經過樣式比對後的特徵詞為 5000 個，過濾後的特徵詞字典則有 542 個特徵詞。



圖三、特徵詞辨識流程

在我們所抓取的評論語料庫當中，將符合樣式（Pattern）以及名詞的數量相除所得到的數值，經過這樣初步的判斷後取出五千個特徵詞，再經由名詞出現次數的排序後，過濾所計算出來的樣式數值，取前一千個大於 0.1、前兩千個大於 0.2、前三千個大於 0.3 的特徵詞，而剩下 542 個特徵詞即是本研究中所使用的特徵詞字典。公式定義如 3-6 所示：

$$FW\ list = \bigcup_{k=1}^{3} Top_{(((k-1)*1000)+1)\sim(k*1000)}\ of\ \frac{\#\ of\ (pattern\ match(N,Adj))}{\#\in N} > k*0.1 \quad (0\text{-}6)$$

N set 包含「*名詞-サ変接続*」與「*名詞-一般*」，*Adj set* 包含「*形容詞-自立*」與「*名詞-形容動詞語幹*」，而樣式部分則給定「*Adj set + feature*」與「*feature + はがも + Adj set*」兩種組合情況。最後於給定特徵詞字典的加權分數上為 0.08，此部分為經由開發資料集（Development set）中的實驗所給定，如公式 3-7。若在該輪計算中的周邊相關字的日文詞有包含在此特徵詞字典內，則在加權值上與不包含的情況做區隔。如：正在計算的相關字是「房子」，此字的日文詞為「部屋」包含在特徵詞字典內。

$$\mathrm{F} = \begin{cases} 0.08,\ s_i\ in\ Japanese\ \in feature\ word\ dictionary \\ 0,\ s_i'\ in\ Japanese\ \notin feature\ word\ dictionary \end{cases} \quad (0\text{-}7)$$

在標註及觀察評論語料後，可發現評論句的特徵詞（評論對象）以名詞為主，評論詞則以形容詞為主。名詞於日語詞性分類中係屬於自立語，因不具活用變化又稱為「體言」，其下可大分為五類：普通名詞、專有名詞、代名詞、數量名詞與形式名詞，另又可依其用法再作細分，如サ行接續名詞、名詞ナ形容詞語幹（亦可稱ナ形容詞）、具連接詞性質名詞、一般名詞等。然而在我們抽取出的資訊中發現，特徵詞的詞性以一般名詞為多數，少部分則為サ行接續名詞（後面可接上サ行變格動詞-する，成為動詞者），一般名詞即如「部屋」（房間）、「絨毯」（地毯）、「ホテル」（旅館）、「スタッフ」（工作人員）等，而屬於少部分的サ行接續名詞則有「食事」（飯菜）、「料理」（菜餚）、「接客」（接待客人）等；另在評論詞方面，前面提過此部分以形容詞主，該詞性的作用為形容或說明事物的性質、狀態，在日語文法上可區分為兩類，一是イ形容詞，另一則是名詞ナ形容詞語幹，イ形容詞後面可以直接接上名詞，如「美しい花束」（漂亮的花），名詞ナ形容詞語幹則必須加上「な」才可以接上名詞，如「静かな場所」（安靜的地方），因此分別歸納出「*N set*」與「*Adj set*」兩類，「*N set*」用來判別評論句中的特徵詞，而「*Adj set*」則是用來判斷評論詞。

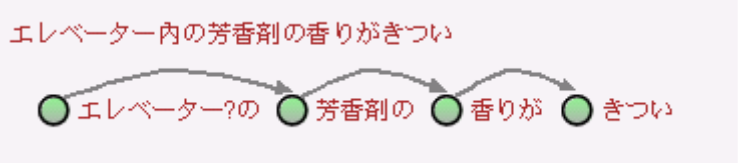
使用者在發表的評論中，特徵詞常會伴隨著評論詞出現，並存在某種語法結構的關係，藉由分析這些結構，就能夠釐清描述評論的特徵詞及評論詞之間的關係，並且歸納出一

些樣式，或稱做模板，而利用這些樣式即可確定何者是我們想要取得的特徵詞或評論詞[13]。在日文語法結構中：(1)「主語-述語」(2)「修飾-被修飾」的關係對於分析評論的特徵詞與評論詞有相當重要的影響，「主語-述語」的關係為構成日文句子的基本結構，主語在句子中為表示性質、狀態的主體以及判斷、說明的對象，是由體言（名詞、代名詞）後面接續助詞所構成，而述語則是敘述主體的性質、狀態以及對某事物的判斷、說明等，是由用言（動詞、形容詞、形容動詞）所構成；另外，當述語的詞性如果為形容詞、形容動詞時，則稱之為描述句（形容詞、形容動詞句），功能為描述主語的性質或狀態，以「風呂は狭い」(浴室很狹小）一句為例，句子中的「風呂」(浴室）為「狭い」（狹小）所要表達其狀態的對象，因此該句的主語為「風呂」，述語則為「狭い」，又因為其詞性為形容詞，故而該句又可稱為描述句。另外，「修飾-被修飾」的關係為形容詞修飾名詞的結構，形容詞屬於用言的一種，也就是具有語尾變化的詞類，依日文語法主要分為形容詞和形容動詞二類，形容詞原形皆以「い」結尾，其後可直接連接欲修飾的名詞，例如形容詞「美味しい」(好吃的)，若要表達「好吃的早餐」即為「美味しい朝食」；而形容動詞的功用與形容詞相同，亦是用以修飾名詞，且也是用言的一種，然而其連接方式卻與形容詞不同，不能直接修飾名詞，當其要修飾名詞時，其後必需加上語尾「な」，例如「綺麗」(乾淨的)，若欲表達「乾淨的房間」為「綺麗な部屋」，故而單獨分為一種詞類。在這邊可以歸納出以下樣式規則：「*Adj set（+な）+feature*」。

然而在使用者發表的句子中，其表達評論的句法不僅限於上述的樣式規則，上述的樣式規則中，評論詞主要會出現在特徵詞前，此屬於前面所提結構中的「修飾-被修飾」句法，另前所提「主語-述語」句法結構亦大量出現於語料之中，除前面所舉的「風呂は狭い」，結構為「主語+は+述語」的句法外，日語語法中連接主語與述語的結構尚有如：「食事が美味しい」(飯菜很美味)，結構為「主語+が+述語」，以及「風景も美しい」（風景也很漂亮)，結構為「主語+も+ 述語」這類的句法，其中助詞「は」與「が」有提示主語的作用，兩者的差別在於所強調的部分不同，助詞「は」強調的是後面述語的部分，以前所舉「風呂」一句為例，其主要強調的地方便是在於「狭い」(狹窄）的部分；而「食事が美味しい」一句中，所強調的則是主語——「食事」(飯菜)，在這邊可以歸納出以下樣式規則：「*feature+はがも+Adj set*」。

（三）、相依性關係

相依性關係（Dependency relation）是經由 CaboCha（南瓜，http://chasen.org/~taku/software/cabocha）所判斷出來的。Cabocha 是以 Support Vector Machines 為基礎的日語修飾解析工具。根據統計自 2001 年 6 月至今日為止，此為日語相依性關係解析工具（Japanese dependency parsing）中準確率最高（89.29%）的系統。因此在效率上較其他工具佳。



圖四、相依性關聯範例

本研究利用此工具取得測試評論句中我們所要判斷的意見詞與其他詞的關聯距離。如圖四中的例句「エレベーター内の芳香剤の香りがきつい」(電梯內芳香劑的香味很強烈)，在此例中我們所要判斷的意見詞為「きつい」(強烈、累人、擁擠)，而利用相依性工具後可以取得「きつい」所屬的節點與其他節點之間的距離，在與「きつい」距離為 1 的節點「香りが」中包含了關鍵字「香り」(香氣)，在與「きつい」距離為 2 的節點「芳香剤の」中包含了關鍵字「芳香」(芳香)，再從其他節點中尋找具有中文翻譯的字來做加權計算。經由公式計算後，此加權值可以使得距離越遠的關鍵字權重越小。計算公式如 3-8：

$$D = 1 + \log(1 + \frac{1}{distance}) \qquad (0\text{-}8)$$

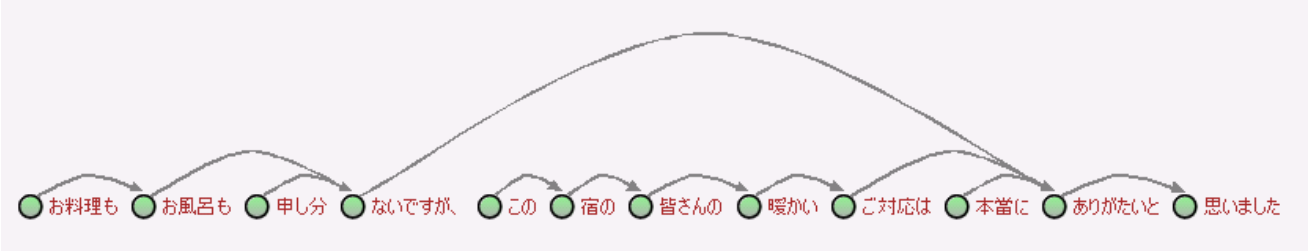
圖四中的例句以另一觀點來看，就是這個句子的重點在最後一個「きつい」，然後每前一個節點在解釋每個後面的節點，如圖五的說明。

～の～の～	→	～的～的～
僕の母の誕生日	→	我的媽媽的生日
○　は（が）△	→	○（東西）△（怎樣）
彼はかっこいい	→	他很帥
～の～の　○　は（が）△	→	エレベーター内の芳香剤の香りがきつい

圖五、相依性關聯分析

另外觀察到，針對日語評論句，我們以每句中的評論詞為基準，觀察句中相關字與評論詞間的距離關係，並對照中譯後過的距離關係，發現在多數評論句中具有「日文距評論詞愈遠，中文距離則遞增」。如以下例句與圖六所示：

- 例句:お料理もお風呂も申し分ないですが、この(3)宿の皆さん(1)の『暖かい』ご対応(1)は本當にありがたい(2)と思い(3) ました
- 中譯：對於飯菜還是浴室都還算令人滿意，但我覺得(思い(3))這間(この(3))旅館所有工作人員(皆さん (1))的『熱情』應對(対応(1))真的讓人非常感謝(ありがたい(2))。



圖六、相依性關聯範例 2

該句評論詞為「暖かい」(熱情)，同句的相關字「皆さん」(各位)、「対応」(應對)分

別在其前後，且距離亦都只隔一個日文假名，故距離同時為 1，然而再看相關字「ありがたい」(感謝)、「この」(這間)、「思い」(覺得)，因為距離評論詞的遠近，距離分別是 2 和 3；我們續而觀察中譯後的句子，並且對照原文相關字於中譯後句子裡的位置，可以發現日文句子於進行中譯時，雖然會有所調整以讓翻譯後的句子符合中文語法，致使原文與中譯的詞語位置或順序產生不同，但是經過翻譯的相關字與評論詞間之距離，仍具有「日文距評論詞愈遠，中文距離則遞增」的現象，如例句的中譯，相關字「皆さん」、「対応」分別翻譯成「所有工作人員」以及「應對」，經過翻譯後，該二字之於評論詞的位置與日文原文相同，分別位於評論詞的前後，其中「所有工作人員」與評論詞間隔 一個中文字，而「應對」則直接皆於評論詞之後，距離次之的相關字「ありがたい」(感謝)於中譯句評論詞的位置相較於另外距離更高的「この」(這間)、「思い」(覺得)要來的 近，其中相關字「思い」 雖因為翻譯關係，所在位置與原文句不同，然距離評論詞仍是較為遙遠。

四、實驗

(一)、實驗資料

本研究主要以日本樂天旅遊網站上的評論資料為主，樂天旅遊網站上規劃住宿評論功能，提供使用者於住宿之後以樂天旅遊會員身份針對所住宿的旅館進行評論，使用者亦可於評論網頁撰寫感想、對旅館的建議、批評等訊息，旅館經營者亦可藉由此平台回覆使用者的批評或建議。

其中我們抓取所有評論當作在辨識特徵詞所用的語料庫，並從中挑選四個常見的日文意見詞進行實驗，主要是在實驗前收集了一些旅遊中常用到的意見詞，而有的詞義並不會差異到影響評論的判斷，這四個字是比較會影響判斷的詞。這四個字的日文詞、詞義、測試句子筆數、平均評論句長度、最長評論句長度、最短評論句長度列於表一中。從表一中可看到測試資料共有 2945 個例句，而句子的平均長度為 30 個日文字長度，最長長度為 70 個日文字，最短則為 5 個日文字。而正確答案的標註是由兩位熟悉日文的專家來標註。

表一、測試資料統計

Word	Word Sense	#	Avg.	Max.	Min.
綺麗	漂亮、美麗 / 乾淨、清潔	647	28	61	8
暖かい	熱呼呼、暖和 / 熱情、親切	1062	31	66	9
きつい	累人 / 強烈 / 擁擠	560	32	73	7
冷たい	冷淡 / 冰涼	676	28	70	5
			30	68	7

（二）、實驗設計

本研究實驗中，主要是以該日文詞的詞義是否判斷到正確的意思為主，因此使用正確率（Accuracy）來作為本系統的評斷標準。而我們系統與 Most Frequent Sense、C. Monz 與 B. J. Dorr[11]、以及知名的線上翻譯 Google（http://translate.google.co.jp）的翻譯結果做比較。Most Frequent Sense（MFS）為測試資料中，統計該測試日文詞多數句子屬於哪一種詞義，則將所有測試句子給予該詞義來作為最簡單實作的基準，亦即測試資料中屬於某詞義數量最多的筆數除於所有測試句子數。

C. Monz 與 B. J. Dorr[11]的方法則完全不經其他修正的實驗情況，除了在實驗時的語言、字典、語料庫的不同外。線上搜尋引擎則以較知名的 Google 翻譯來當作另一種評分基準。

（三）、實驗結果

下表二為本研究的實驗結果，其中 MFS、Google、Iterative(I)為本實驗的基準組。Google 翻譯是經由兩位標註者經由整句翻譯查詢結果後，僅針對本研究中所要翻譯的意見詞來做判斷，而不受其他字詞翻譯錯誤的影響；Iterative(I)則為 C. Monz 與 B. J. Dorr[11]中研究方法的一部分，在表格中以 I 代表該方法。Abs 則代表在第三章所敘述的在正規化步驟時加上處理絕對值的情況，因原方法 I 僅處理正規化步驟，並未加上絕對值處理。F 與 D 分別代表第三章所介紹到的特徵詞字典加權與相依性關係加權。

表二、實驗結果

	綺麗	暖かい	きつい	冷たい
MFS	70.47%	50.28%	53.57%	65.97%
Google	62.90%	57.43%	13.03%	61.68%
Iterative (I)	44.82%	42.74%	17.50%	15.23%
I + Abs	75.88%	56.30%	52.85%	80.62%
I + Abs + F	77.89%	59.51%	54.82%	82.39%
I + Abs + D	79.59%	60.73%	56.96%	84.76%
I + Abs + F + D	81.45%	62.05%	57.67%	86.24%

由表二的實驗結果可觀察到，在線上翻譯方面，Google 比 MFS 的效果來得差些。若以整句的意思與或翻譯後的中文句來比較，Google 翻譯的結果反而令人難以了解。

而從 Iterative 的結果來觀察，可以發現在正規化的過程中是否有處理絕對值的部分影響頗大，這部分可歸因於所使用的語料庫不同的關係，C. Monz 與 B. J. Dorr[11]的研究中所使用的語料庫為 CLEF 2003（Cross-Language Evaluation Forum）英語與德語的雙語資料，因此在語料庫的品質上會較單純從網路上所構成的語料庫來的好，由網路上無論是片段摘要或者整頁的文字資訊多多少少會有雜訊，較無法完全過濾乾淨，尤以片段摘要較為嚴重。

其他在增加特徵詞加權以及相依性關係詞加權後皆有改善整體的正確率。而加上特徵詞加權以及相依性關聯詞加權的分數為最高，平均約比單純修正絕對值多了 6%的正確率；比線上 Google 翻譯多了 23%；比 MFS 的基準多了 12%。

五、討論

在本章節，我們描述對於日語意見詞翻譯結果後，容易造成系統判斷錯誤的一些問題。

（一）、斷詞錯誤

在例句「エントランスはびっくりするほど綺麗」中，有出現斷詞錯誤的情況，因而造成線上字典翻譯也跟著錯誤，例如「エントランス」(entrance)原本的意思應為「入口」，而在日文斷詞的時候已經將此詞斷為「エン」與「トランス」兩個詞，因而在查詢線上字典翻譯時查找到分別對應二者的中文翻譯「圈子、圓圈、日元」與「變壓器」，在這邊意思已與原先的不同，進而造成在後來抓取網頁的片段摘要（Snippet）以及計算 Mutual Information 的時候的錯誤。正確的 Mutual Information 組合「漂亮 AND 入口」原應為有正向關係的情況，而斷詞錯誤過後的組合「漂亮 AND 日元」與「漂亮 AND 變壓器」則為「無共現情況」與「-0.5」的數值。

另外，有些斷詞錯誤的情況我們將其另外分區隔，最明顯的差別在於其必須同時符合下面兩個條件：一是於字典上無查詢結果，二是會被斷詞工具斷開成兩個字，例如：「ソファーベッド」(沙發床)應該是一個詞，然而於線上字典（Excite 日中）無查詢結果，且經過斷詞後會成為「ソファー」(沙發）與「ベット」(床)，再如「ダブルベッド」（雙人床)，於線上字典（Excite 日中）亦無查詢結果，且其斷詞結果亦被斷開為「ダブル」(重、重複）以及「ベッド」(床)；再把斷詞後的錯誤結果於字典進行查詢，有些可得到查詢結果，然而有些卻無法得到。

會出現此情況的以片假名居多，由於歸入此類者在尚未進行斷詞前便已無法於字典中得到查詢結果，此與字典所收錄的詞條資料有很大的關係，另外部分片假名因為組合的關係，其斷詞錯誤的結果可能各自具有其意思，即便再於字典中查詢，但實際上在進行抓取網頁片段摘要的時候卻已無法取得最正確的意思，因而造成在計算上的錯誤。

（二）、片假名情況

在日語中以片假名書寫的文字除了發表者為了強調該詞語之外，外來語還是片假名書寫的大宗。目前日語外來語大致有下列幾種形式：一、音譯，如「ガラス」(玻璃)為「glass」；二、外語略稱，如「ビル」(大樓）本應表記為「ビルディング」(building)；三、和製外來語，如 オートバイ（摩托車）來自於「Auto」與「Bike」。再者，日本國內轉化外語製造新語的速度十分驚人，幾乎每天都有新造語出現，不僅讓學習日語的外國人感到頭痛，就連日本國人對這種情況感到困擾者也不在少數，因此也就容易造成字典無法檢索出某些詞彙的現象；除此之外，也與字典建入新詞條速度的快慢有關係，一般辭典並不會更新如此快速也是造成此問題的原因之一。例如例句「建物や部屋は新しく綺麗でとてもスタイリッシュで格好良い」中的「スタイリッシュ」正確意思為「時尚」，而在系統中則因為於線上字典中查找不到相關翻譯，而遺漏了對於我們所要判斷的意見詞

「綺麗」有幫助的周邊相關字資訊。

（三）、中文用詞不同

本研究所使用的 Excite 翻譯雖然提供中日、日中的辭典查詢，但是該系統的輸出主要是簡體中文，目前網路上可見的中日／日中辭典查詢結果皆為簡體中文，今日華語使用者圈中所使用的華語可大別為簡中與繁中兩類，用語除寫法有所差別外，在語意以及詞語使用上雖大致上相同，但由於語言會受到使用者所生長的環境、文化、風俗以及習慣不同，而多少衍生出差異，以今日中國大陸與台灣的用語為例，雖大致一樣，然而仍可以發現於指涉某些事物上其用語仍存在著差異，例如大陸台商經貿網（http://www.chinabiz.org.tw）上所公佈的兩岸常用詞語對照表中，「人工流產」在台灣稱為「墮胎」，中國則稱為「人流」；再如「馬鈴薯」是台灣說法，中國卻稱「土豆」;「調理包」在中國則稱為「方便菜」，上述這些現象也多少出現於所查詢的日中辭典中，查詢結果會有少數候選詞組會出現台灣人較少使用的用法，例如評論資料中「おしぼり」一詞，線上辭典翻譯為「手巾把兒」，對於台灣的華語使用者來說並無法在一見到該詞的當下便立即瞭解意思，而《新時代辭典》對該詞的翻譯為「濕手巾」[14]；再如「キビキビ」一詞，線上辭典的翻譯是「脆、俐落、俏皮、爽利、麻利」，其中「麻利」係屬於中國陝西方言，於台灣幾乎不曾使用，然而《新時代辭典》翻譯為「機敏、爽快」[14]，相較於線上辭典翻譯來說要來得親切。而上述的情形也造成在擷取繁體中文網頁摘要片段的時候，資訊會較其他繁體中文中多數用法的次數來得要少。

（四）、翻譯錯誤

這部分說明有些被判斷錯誤的原因之一，為翻譯錯誤。若在測試句子中前後的相關字有出現翻譯錯誤的情況，會造成在利用網路資源製作語料庫時會抓取到錯誤的資料，進而影響在計算判斷過程中的機率，導致在最終的詞義判斷上出錯。我們在此觀察列出五種原因：

1、同音不同義

同音不同義所指的是所抓取到的結果並非原本搜尋字詞的正確意思，而是與其同音然而意義卻不同的字詞。如例句「朝食の内容ですが、個人的には朝から丼物はきつい感じで、やはりスタンダードな洋食と和食の組み合わせのほうが良いと思います」的「ほう」一字，按所抓取到的字典搜尋結果為「法律」，然而若以該結果對照其於例句中發表者所意欲表達的意思，則明顯發現發現意義產生出入，原因在於「ほう」一字在該例句的意思應當為「方面」，而非「法律」。

2、日語中的用言

所謂「用言」，乃是指日文中具有活用型的詞，這些詞語會隨著時態（過去或現在），或者是肯定與否定，抑或是對話者身份（敬體、常體）而有不同的活用型，以「美味しい」（好吃）一詞為例，假若說話者要表達是過去時空中吃到的東西很好吃，那麼該詞便會在語尾產生變化，成為過去式「美味しかった」，這本是日語的特性之一，然而卻也相對地造成搜尋字詞翻譯時發生謬誤的情況，以「観光地なのに対応が冷たいホテルもあります」（明明是觀光區但還是有待客態度冷淡的旅館）一句為例，句中的「あります」

係日語中的丁寧體，原型為「ある」，中譯為「有、具有」，以人工來檢視並不會造成字義上的謬誤，然而經過斷詞之後會被斷為「あり」與「ます」兩個部分，繼而查詢詞典後，「あり」原本的中譯應為「有、具有」，其在字典的查詢結果卻是「螞蟻」。

3、字典未收錄

Excite 日中字典中雖收有三萬詞條，但是仍有尚未收錄的字詞翻譯，如「寛げる」（舒暢）、「和室」（和室）、「館内」（館內），於字典的查詢結果皆是「キーワードに該當する結果が見つかりませんでした」（未有所查詢關鍵字之結果）。

4、形同義異

形同義異，亦即字詞的寫法相同但意思卻完全不同。日本人於轉化歐美文字製作新語方面，亦有其習慣方式，例如英文字母「A」可被標記為日文假名「ア」，在單字轉製方面，以英文字「best」為例，「be」標記為「ベ」（be），「s」則為「ス」（su），「t」則以「ト」（to）標記，而成為「ベスト」，此種讀音轉化方式雖能使日本能可以快速吸收外來文化或新知，但卻也容易產生字詞假名標記相同、字義卻無任何關連的情況，假如再以「ベスト」一字查詢字典，便會發現其除了「最好」之外還有同樣寫作「ベスト」但意思卻是「背心」的查詢結果，且字源與「最好」來自於 best 不同，而是來自於「vest」一詞；而以語料中所見字詞「バス」為例，根據日文字典中查詢該詞的結果則有：bass（男低音）、bath（浴室）、bus（公車），因此可能對應到的翻譯便有上述三個；且此情形不惟出現於外來語，在日語原有的詞彙上亦會產生相同的情況，如「辛い」一字，讀音便有「karai」與「tsurai」兩種，但是意思卻全然不同，「karai」意為「辣」，「tsurai」則是「難過」的意思。

六、結論

本研究著重耗費成本極小的非監督式導向詞義消歧方法，並將方法應用在日語旅遊意見詞翻譯上，讓使用者在瀏覽閱讀評論時能更快速決定是否要將其納入考量之一，並去除了語言上的障礙，不會因語言不通而僅能取得資訊較少的中文評論。本研究所提出之方法包含語料庫準備、特徵詞字典產生、語法相依樹以及加權方式，以決定意見詞翻譯候選詞，幫助使用者更加了解評論之內容。由實驗結果顯示，我們所提出的方法對於翻譯日語意見詞是相當有效的，其結果皆比三組基準值包含常用的線上翻譯系統的正確率要來的好。透過本篇論文所提出之方法，使用者將能夠更清楚的理解評論的意義，且不會因為線上整句式的翻譯不通順造成無法理解評論所想要傳達的原意。

在未來可搭配意見分析領域的特徵詞與意見詞配對，例如：「服務 – 熱情」、「房子 – 漂亮」等配對，對於華人在自助旅遊快速收集與整理所要的資訊，必能節省相當大的時間成本。

參考文獻

[1] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, 2007.

[2] C. Cardie and R. J. Mooney, "Guest Editors' Introduction: Machine Learning and

Natural Language," *Machine Learning*, vol. 34, pp. 5-9, 1999.

[3] D. McCarthy, "Word Sense Disambiguation: An Overview," *Language and Linguistics Compass*, pp. 537-558, 2009.

[4] U. S. Kohomban and W. S. Lee, "Learning semantic classes for word sense disambiguation," presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan, 2005.

[5] R. Florian, *et al.*, "Combining Classifiers for word sense disambiguation," *Natural Language Engineering*, vol. 8, pp. 327-341, 2002.

[6] Z.-Y. Niu, *et al.*, "Optimizing feature set for chinese word sense disambiguation," presented at the Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems, 2004.

[7] W. A. Gale, *et al.*, "A method for disambiguating word senses in a large corpus " *Computers and the Humanities*, pp. 415-439, 1992.

[8] D. McCarthy, *et al.*, "Finding predominant word senses in untagged text," presented at the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain, 2004.

[9] 梅家駒, *et al.*, *同義詞詞林*: 東華書局, 1993.

[10] Z. Lu, *et al.*, "An equivalent pseudoword solution to Chinese word sense disambiguation," presented at the Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Sydney, Australia, 2006.

[11] C. Monz and B. J. Dorr, "Iterative translation disambiguation for cross-language information retrieval," presented at the Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil, 2005.

[12] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, pp. 22-29, 1990.

[13] N. Kobayashi, *et al.*, "Collecting evaluative expressions for opinion extraction," presented at the Proceedings of International Joint Conference on Natural Language Processing, 2004.

[14] 陳伯陶, *新時代日漢辭典*: 大新書局, 2005.