The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 281-295

基於語境特徵及分群模型之中文多義詞消歧

Using Contextual Information in Clustering Methods for Chinese Word Disambiguation

李右元 周子皓 劉昭麟

Yu-Yuan Lee, Tzu-Hao Chou, Chao-Lin Liu

國立政治大學資訊科學系

Department of Computer Science

National Chengchi University

{107753027, 104753029, chaolin}@nccu.edu.tw

摘要

多義詞是語言中極為常見的現象，在過去，若要查找多義詞的義項及其使用方式，必須翻查傳統辭典，但礙於篇幅問題並非所有義項都會收錄，因此所提供的例句數目也較少。即使隨著科技進步，發展了數位化的辭典與檢索系統，仍有部分問題存在。因此，人文學者必須耗費大量心力以人工判讀方式辨別義項的不同。

本研究以分群模型將大量已向量化之中文語料加以處理，透過 purity 分數比較出最適之模型，並挑選適量的例句供使用者參考。實驗中以人工標記之例句作為評分依據，結果顯示屬於同形異義(homonymy)之多義詞在 macro-average、weighted-average 與 accuracy 皆能達到 0.85 以上之水準。

Abstract

We present preliminary results for searching for useful sentences for learning ambiguous words with clustering methods. First, we search for sentences that contain an ambiguous word (the target word, henceforth). To make the extracted sentences useful for learning the target word, we attempt to guide the clustering methods to separate the sentences that carry different senses of the target word into different clusters. We influence the functioning of a clustering method

by providing example sentences that carry specific senses of the target word. In the terminology of machine learning technology, we label a sentence with the sense of the target word in the sentence. Two sample labeled sentences for the ambiguous word “bank” follow.

1. “financial institution”: Mr. Black deposit the money in the Citi bank.
2. “place”: Along the bank of the Charles river, you may see the MIT campus.

Assume that we can collect a large number of sentences that contain the target word, for which we need sentences that use a specific sense of the target word. Assume that we are willing to label a few of these original sentences as we described above.

A clustering algorithm may employ the labeled sentences to build clusters of sentences for our needs. The algorithm may take advantage of the labeled sentences as informative seeds for initializing the clusters. In addition, when selecting the (unlabeled) sentences from the clustered sentences as the final output, the labeled sentences may also provide guidance for selecting the sentences of “correct” senses. If a cluster has many labeled sentences of a specific sense, the (unlabeled) sentences in this cluster might have the same label of the sample sentences. Furthermore, to select and output the (unlabeled) sentences in this cluster, we may consider the (unlabeled) sentences that are closer to the sample sentences.

Assume that we may find thousands of sentences that use a target word, assume that we provide a certain number of labeled sentences to guide a clustering algorithm, assume that we cluster the thousands of sentences into tens of clusters, and assume that we select just tens of sentences from these tens of clusters. If our clustering methods are good and if we select sentences from a cluster conservatively, we may achieve high precision in the final selection of the unlabeled sentences for the target word.

Empirical evaluations reported in this paper show promising results. Not surprisingly, we found that it was relatively easier to achieve better results for homonym than for polysemy. We hope our methods can be useful for building corpora for learning ambiguous words.

關鍵詞：多義詞，一詞多義，同形異義，分群模型，詞向量，句向量

Keywords: lexical ambiguity, polysemy, homonymy, clustering, word vector, sentence vector

一、 緒論

多義詞存在於大多數的語言當中，Bruce Britton[1]曾保守估計至少 32%的英文單詞存在著不只一個義項，且最常使用的前一百名英文單詞中，高達 93%的單詞是多義詞，可見多義詞與人類的生活是息息相關。此外，多義詞還能細分成一詞多義(polysemy)與同形異義(homonymy)[2]，前者定義為其義項類別彼此有關聯，後者則是彼此無關聯。以 “cup” 為例，茶杯的 “cup” 與獎盃的 “cup” 是一詞多義，而銀行的 “bank” 與河岸的 “bank” 則是同形異義。

以往若要瞭解詞彙的其他義項，必須透過辭典進行查找，但是辭典中提及的義項大多是較有規範化的使用方式，其內容較少包含口語化的義項類別，且礙於篇幅關係，並非所有義項都能有足夠的例句讓使用者瞭解該義項的使用方式。此外，編輯辭典耗時費力，實在難以符合語言發展的速度。

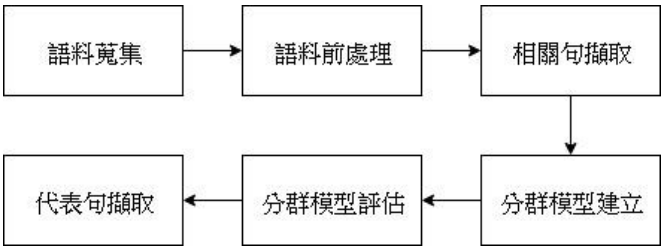
隨著網路與資料庫的發展，有許多的數位檢索系統扮演著新型態的辭典供使用者查詢。以「教育部重編國語辭典修訂本」[3]及「萌典」[4]為例，藉由線上檢索的功能或許能改善部分傳統辭典的問題，但仍觀察到某些較新興的詞彙，如「亞馬遜」不存在於資料庫內，亦或者例句數量較少，難以讓使用者清楚明白適用該義項之句型。再以「中央研究院現代漢語標記語料庫」[5]為例，其搜尋結果沒有顯示出義項類別，因此仍必須透過人工方式逐一閱讀。

過去曾有人文學者對於特定詞彙以人工方式，透過其在特定領域的專業知識，判別出過往辭典中未收錄的義項類別，例如吳美嫻[6]、林香薇[7]、蔡宛玲[8]與許尤芬[9]等人的相關研究。在資訊領域，Navigli[10]曾提出對於消除歧義有監督式、非監督式及以知識為本等三種實驗方式，而 Lesk[11]為首位提出以知識為本的消歧方式，透過計算目標詞彙所在上下文與目標詞彙在知識庫中各義項類別的覆蓋程度，對目標詞彙進行消歧。

因此本研究嘗試透過文字向量化的方式，將大量的中文語料與少數使用者提供之參考句轉換成向量，再以分群模型演算法將相同義項類別之相關句進行區隔，同時搭配 purity[12]分數尋找出最適合該多義詞之分群模型，最後以實驗中設計的方法擷取出一定數量的代表句。透過分群模型，提供例句供使用者閱讀。

二、研究方法

在本研究中，主要分為六個階段，分別如圖一所示。



圖一、研究方法流程圖

（一）、語料搜集

本研究所使用的語料主要是中文維基百科所提供的開放資料[13]、教育部 AI CUP 2019 新聞立場檢索技術第一階段所提供的新聞資料[14]，以及由使用者自行提供的參考句。

維基百科是個開放式的協同寫作平台，其內容是由眾多使用者與專業人士一同撰寫而成，因此不會受限於個人或團體的寫作風格與觀點，其收錄內容不僅包含傳統百科所蒐集的資訊，由於線上協同寫作的關係，其內容更容易與時俱進收錄到較新的詞條訊息。

然而，維基百科雖然是開放式協同合作的資料，但其終究是屬於百科類文體，受限於此，可能導致我們所在意的目標詞彙的義項比例分布不均勻。因此，在實驗中我們加入了約十萬篇的新聞語料藉以平衡這樣的問題。新聞語料內容來自中國時報、自由時報、蘋果日報、聯合報與 TVBS 等媒體，內容橫跨各種議題，包含政治、經濟、教育、…等等，用以增加語料的多元性。

此外，研究中亦會請使用者提供具有人工標記義項類別之句子，不僅為語料增加更多的文字風格外，也能作為後續實驗方法的依據之一。

（二）、語料前處理

不論是維基百科或者是新聞語料，都夾雜了或多或少的額外訊息，因此在進行實驗前必須費工處理以得到乾淨的文本。另外，由於中文的書寫結構與英文不同，因此通常需要額外的方式來加以處理。

以維基百科為例，我們首先使用 WikiExtractor[15]從原本取得之 XML 格式的資料內容中過濾掉大量不需要的資訊，僅留下條目標題、條目超連結、條目編號及條目正文等資訊，接著透過 OpenCC[16]將原本繁簡混雜的內容統一轉換成臺灣繁體，舉例來說，OpenCC 會將原始簡體內容的「光盘」轉換成臺灣使用的「光碟」。最後再去除剩餘不需

要的標籤與資訊，最終得到我們所需要的內容。

將所有語料都處理乾淨後，我們首先必須做斷句的工作，研究中是以逗號、句號、驚嘆號與分號這四種標點符號做為句子之間的斷點，我們將斷開後的句子稱為例句，表一說明實驗中使用的維基百科與新聞語料的數據統計。

表一、語料數據統計

項目	中文維基百科	新聞語料
條目總數	1,092,447 篇	98,250 篇
例句總數	26,815,431 句	5,489,253 句
例句平均長度	15 字	15 字
條目平均例句數	25 句	56 句

在斷句之後必須處理中文的斷詞問題，因為中文不像英文是以空白做為詞與詞之間的分隔，因此必須依賴斷詞器達到詞彙分割的目的。實驗中我們比對了中研院 CKIP 中文斷詞系統[17]與 Jieba 斷詞器繁體版[18]兩者間的差別，從維基百科中隨機抽樣一百句例句並透過人工斷詞作為標準答案，透過萊文斯坦距離(Levenshtein distance)計算兩者與標準答案間的相似度，最終 Jieba 得到較高的相似度，因此我們採用 Jieba 斷詞器繁體版來處理這樣的問題，藉由 Jieba 能夠有效率地將個別詞彙分開，以利後續工作進行。

（三）、相關句擷取

在本研究中，目標詞彙指的是具有多重義項類別的詞彙，例如「亞馬遜」可以代表亞馬遜雨林，也可以代表電商亞馬遜公司。而相關句是指包含目標詞彙的例句。此外，在現實情況下，一句例句所提供的語境可能無法準確地判斷出目標詞彙在此例句中所代表的義項，因此我們將相關句的左右例句也加入相關句中，亦即一筆相關句中包含三句例句，希望藉由增加相關句的長度藉以提供更豐富的語境資訊。然而因為原始語料標點符號的錯用，可能導致即使已經增加相關句長度，其相關句仍然過短無法提供足夠的語境資訊的問題，因此我們根據相關句中例句的長度來進行篩選，將語境相對貧乏的相關句剔除。

（四）、分群模型建立

在建立分群模型之前，必須將處理好的語料轉換成向量的形式，實驗中使用的向量化模型如表二所示，其中包含四種模型，各模型皆有七種向量維度及七種 window size 作為參數調整。

表二、向量化模型表

Embeddings	Models	Vector size	Window size
doc2vec[19]	PV-DBOW	5, 10, 20, 50, 100, 200, 500	2, 5, 10, 15, 20, 30, 50
	PV-DM		
average word2vec[20]	CBOW		
	skip-gram		

而分群模型則使用了 K-means[21]、hierarchical clustering[22]、spectral clustering[23][24][25][26]與 BIRCH clustering[27]等方法，而分群數量則有[2, 10]等 9 種組合。

其中，起始點的選擇方式對於 K-means 的分群結果會有重要的影響，若起始點的選擇與語料數據的分布情形差異過大，會造成效果較差的分群結果。在實驗中，我們採用了三種不同的起始點選擇方式，以下簡略說明：

方法 1： 以 K-means++[28]作為起始點選擇。有別於原始的 K-means 演算法，起始點是隨機選取的，K-means++會選擇離當前起始點最遠的點作為新群集的起始點。

方法 2： 先將參考句分群，再以分群後的群集中心作為相關句分群的起始點。主要是將由維基百科與新聞語料組成的相關句與使用者提供的參考句分別進行分群工作，因為參考句是由使用者提供，能夠掌控目標詞彙不同義項的例句數量，降低因為例句比例差異造成分群錯誤的問題。

方法 3： 依據使用者提供之參考句的義項類別將參考句分群，再以群集中心做為起點。不同於方法 2 是將參考句全部一起進行分群，而是根據人工標記之義項類別各自對參考句進行分群，再將各自分群的群集中心合併，作為相關句的分群起始點。此外，由於是根據義項類別做分群，實驗中的分群數目設置為義項類別的倍數。以目標詞彙「亞馬遜」為例，其包含雨林與電商兩種義項，則分群數目設值為 2 的倍數，即 2、4、6、…等。

此外，hierarchical clustering 的距離計算分別採用了 ward's linkage、complete linkage、average linkage 與 single linkage 等四種方式。

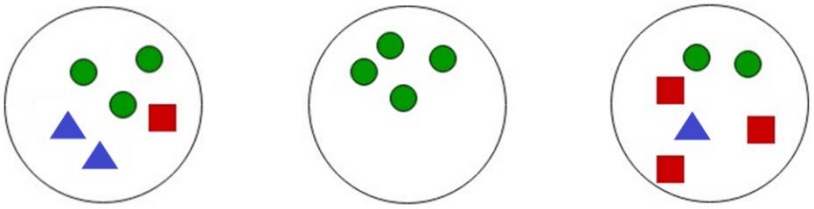
（五）、分群模型評估

如同第四小節所述，本實驗主要採用分群模型判別相關句目標詞彙的義項，並採用了多種分群模型與不同的參數設置模型，目的是要找到一個相對好的模型。實驗中我們採用 purity 作為評估指標，主要是以使用者提供的參考句計算 purity。因為參考句包含人工標記的義項類別，能夠讓我們判斷該群集是以目標詞彙的何種義項為大宗。

若一個群集內的資料其實際類別彼此相同，則 purity 分數高，反之則低，而分數區間是介於 0 到 1 之間。purity 公式如(1)所示，其中 N 代表資料點總數，$\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ 代表 k 個群集，$C = \{c_1, c_2, \dots, c_j\}$ 代表 j 個類別。

$$purity(\Omega, C) = \frac{1}{N}\sum_{k} \max_{j} \left|\omega_k \cap c_j\right| \quad (1)$$

以圖二作為舉例，假設分群模型將實驗語料分為三個群集，圓形資料點為維基百科與新聞語料所組成的目標詞彙相關句，正方形與三角形資料點為不同義項類別的使用者提供參考句。圖二-(a)表示該群集內存在三筆相關句、兩筆屬於三角形義項之參考句及一筆屬於正方形義項之參考句。根據 purity 的定義會加總群集中相對多數的資料點，又因為我們是針對參考句去計算 purity，圓形所代表的相關句不會影響分數的計算，也就是圖二-(b)對於整體 purity 分數不會有影響，僅計算圖二-(a)與圖二-(c)中所佔比例最多之參考句數目。因此範例的 purity 分數為 $\frac{1}{7}(2+3) \approx 0.714$。



(a) (b) (c)

圖二、purity 於本研究中的使用範例

（六）、代表句擷取

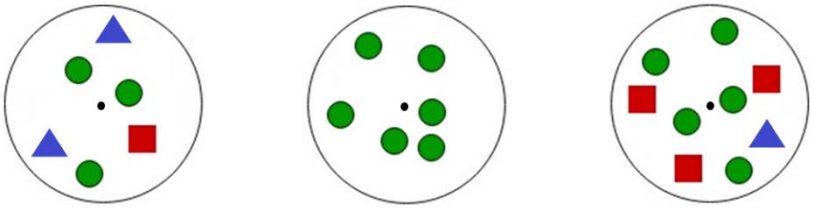
透過分群模型的建立，我們能夠將向量化的相關句分成若干群集，各群集中目標詞彙所代表的義項也不完全相同。為了達到實驗目的，讓使用者瞭解目標詞彙在不同義項上實際使用情形，我們在此定義了代表句為能用以表示目標詞彙使用情形之相關句。並設計了三種代表句擷取方式，再以人工標記之正確答案來評估代表句擷取效果。以下簡略敘

述，其中圓形資料點皆代表相關句，正方形與三角形資料點則為不同義項類別之參考句。

方法 1： 擷取包含參考句在內之群集中的所有相關句做為代表句。以圖三做為舉例，圖三-(a)群集會擷取三筆相關句做為代表句；圖三-(b)群集因為不包含參考句，不做擷取；圖三-(c)群集則擷取五筆相關句做為代表句。

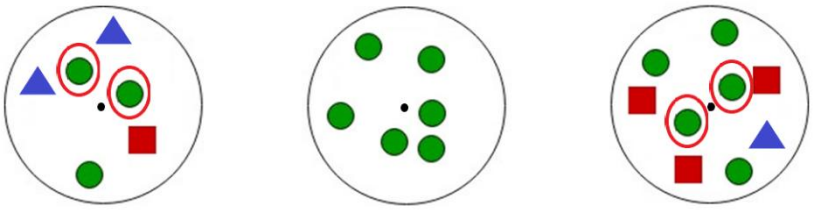
方法 2： 依據相關句與群集中心距離，擷取包含參考句在內之群集中的相關句。以圖四為例，黑點代表群集中心，距離計算採取 cosine similarity 方式，擷取與群集中心距離最近之前兩句相關句做為代表句，其餘距離較遠之相關句則不採納。

方法 3： 依據相關句與參考句之群集中心距離，擷取包含參考句在內之群集中的相關句。以圖五為例，黃色星形代表各群集中參考句之群集中心，同樣採用 cosine similarity 方式計算距離，並擷取與該中心距離最近之前兩句相關句做為代表句。



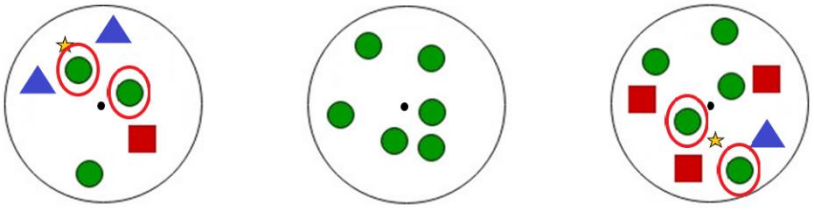
(a) (b) (c)

圖三、代表句擷取方法 1 示意圖



(a) (b) (c)

圖四、代表句擷取方法 2 示意圖



(a) (b) (c)

圖五、代表句擷取方法 3 示意圖

此外，因為向量化具隨機性，為了觀察在不同參數下分群的效果，考量機器效能會以相同參數重複執行十次，透過 hierarchical clustering 將這十次產生的群集視為集合，再度分群。而 hierarchical clustering 的距離計算改為使用 Jaccard Index，而非歐氏距離，其分數越高代表兩群集間相似度越高。最後透過人工標記之答案計算 macro-average[29]、weighted-average 與 accuracy 作為代表句擷取效果評估。

在執行 hierarchical clustering 後可能會有某些相關句同時出現於不同群集中，本研究中設計兩種方式來解決，以下簡略說明。

方法 1： 直接剔除重複出現的相關句。若相關句於不同群集中出現，可能是位於群集邊際所造成，可能與群集內其他相關句的同質性較低，因此參考價值較低，可以直剔除。

方法 2： 計算重複出現之相關句與其群集中心之距離，並將其歸屬於距離較近之群集。為避免直接剔除該相關句可能損失特定資訊，透過距離的比較保留相關句內容。

以下說明代表句合併過程，若資料集內有 {1, 2, 3, 4, 5, 6} 等六筆相關句，將資料集劃分為兩群，重複執行兩次，第一次分群結果為 (1, 2, 3)、(4, 5, 6)，第二次分群結果為 (1, 2, 3, 4)、(5, 6)，再依上述 hierarchical clustering 分成兩群，分別為 [(1, 2, 3), (1, 2, 3, 4)]、[(4, 5, 6), (5, 6)]。此時相關句 4 同時於兩個群集中出現，再依上述方法處理。

三、研究結果分析

（一）、目標詞彙篩選

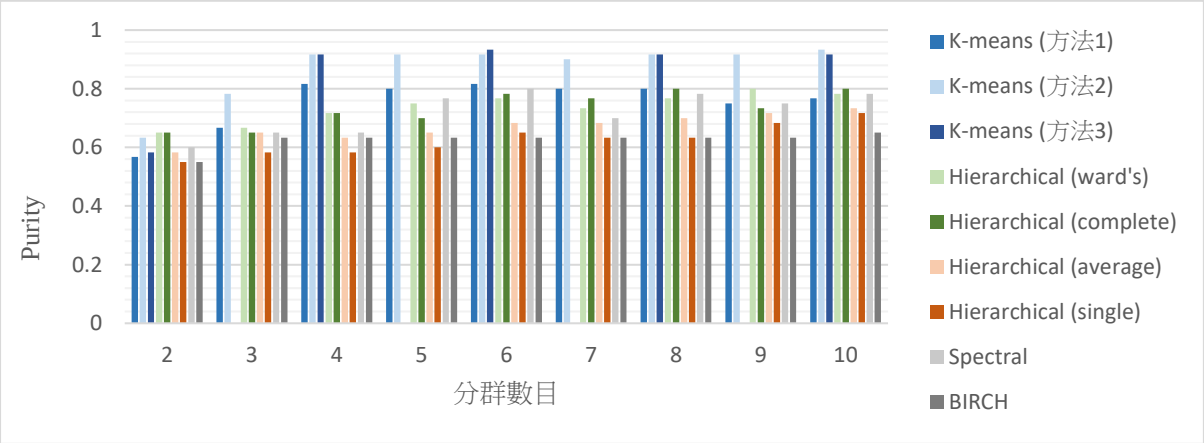
在實驗中，目標詞彙的相關句數量必須足夠，否則機器無法充分學習其語意特徵。此外，因為多義詞有同形異義(homonymy) 與一詞多義(polysemy)的不同，本研究中，同形異義的目標詞彙有「亞馬遜」、「蘋果」、「小米」；一詞多義的目標詞彙則有「出發」、「出入」、「壓力」與「東西」等。表三簡單列出其例句。

表三、目標詞彙義項類別與例句

目標詞彙	義項類別	例句
亞馬遜 (homonymy)	雨林	巴西國家太空研究所氣候學家諾布雷（Carlos Nobre）指出，亞馬遜樹種的死亡和巴西南部
	電商	2015 年黯然退出手機市場。但近日亞馬遜一名高層人員透露，亞馬遜不排除會重返智慧型手機市場
出發 (polysemy)	實際離開	鐵馬進香活動昨清晨 6 點熱鬧展開，300 輛自行車從天后宮出發往北港，挑戰來回 345 公里進香之旅
	從某方面著手	旁觀者到思考者，轉換角色，從整理自己開始，陪考也有重新出發的可能

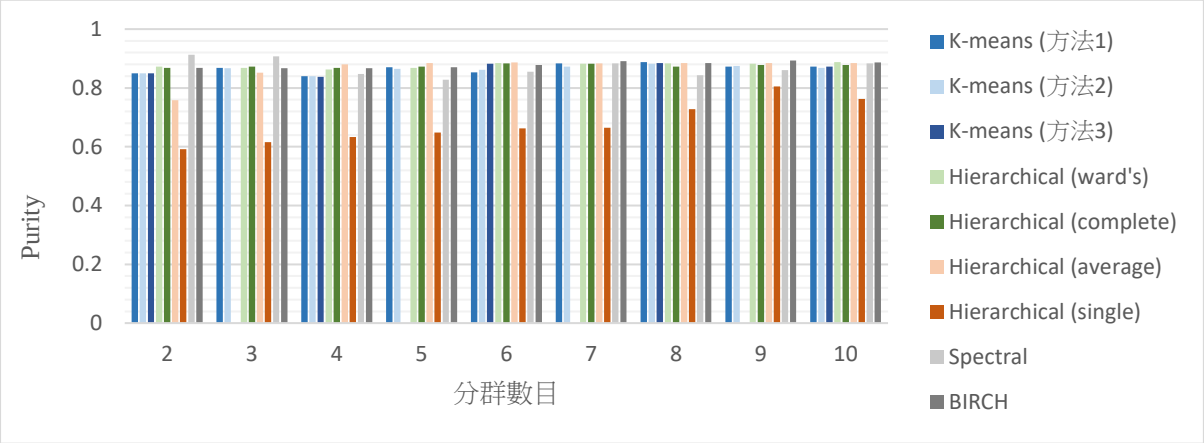
（二）、分群模型評估

在九種分群模型下，分群的數量會對於分群結果造成影響。此外，因為向量化具有隨機性，為了評估分群的效果，實驗中以相同參數重複執行十次並取平均作為該參數下分群的 purity 分數。圖六以「亞馬遜」為例，呈現九種分群模型在分群數目區間為[2, 10]之間，各分群模型效果最優之對應 purity 值。從圖六中可以觀察到，在 K-means 的三種起始點選擇方式下，比起方法 1 單純使用 K-means++的方式，方法 2 與方法 3 能夠獲得較高的 purity 分數。而在 hierarchical clustering 的四種距離計算方法中，ward's linkage 與 complete linkage 都能較另外兩種得到更高的 purity 分數。以實驗結果而言，仍然是 K-means 的方法 2 與方法 3 能夠有效地將相關句進行正確的分群。



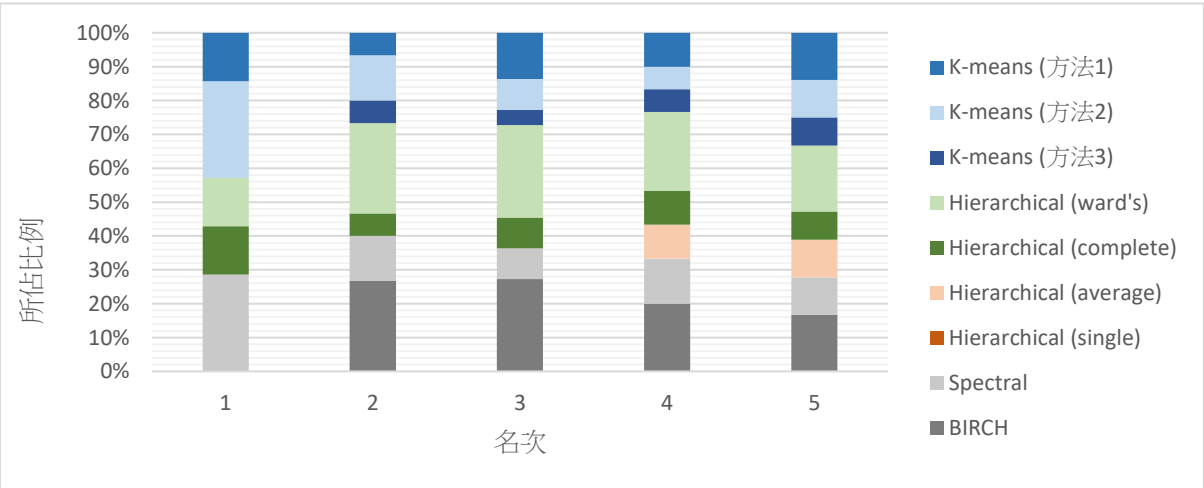
圖六、各分群模型下最優之 purity 分數（亞馬遜）

圖七則是以目標詞彙「出發」為例，以相同的方式重複執行十次，取平均作為 purity 分數。從圖中可以觀察到各分群模型在不同的分群數目下，除了以 single 距離計算方式形成的 hierarchical clustering 外，其餘模型並無明顯變化。根據圖二，可以推測隨著分群數目增加，有許多的群集中可能不包含參考句在內，導致其對 purity 的影響極小。



圖七、各分群模型下最優之 purity 分數（出發）

圖八是依據各目標詞彙在各種分群模型下之最優 purity 分數統計所繪出之圖形，可以觀察到使用 K-means(方法 2) 及 spectral 等分群模型在各目標詞彙最優之 purity 分數出現較多次。而若以各目標詞彙的前 5 名 purity 分數來觀察，可以發現使用 hierarchical (ward's linkage) 方法與 BIRCH 等分群模型所佔比例較高。因此若使用者要以實驗以外之多義詞進行查詢，會以此二種分群模型作為優先推薦。



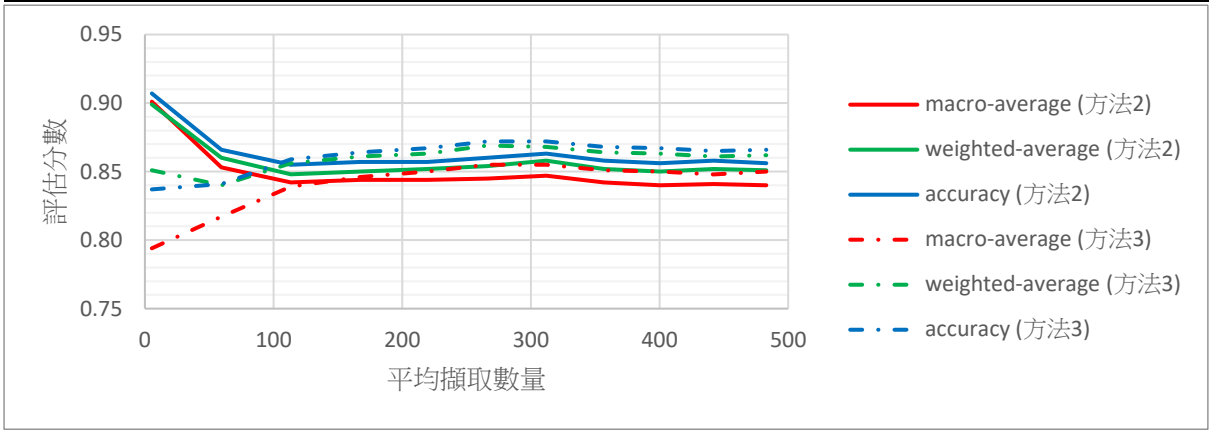
圖八、各分群模型在不同 purity 分數名次所佔之數量比例

（三）、代表句擷取評估

在代表句的擷取，我們採用了三種擷取方式，並透過人工標記的正確答案計算代表句之 macro-average、weighted-average 與 accuracy，作為代表句擷取效果的好壞評估。以下將以「亞馬遜」作為代表，分析在最佳 purity 之參數下，三種擷取方式所得之分數。表四是以擷取代表句方法 1 之參數為實驗，可見三種分數介於[0.840, 0.860]之間。圖九則是以方法 2 與方法 3 之參數為實驗，其中實線為方法 2 之結果，虛線為方法 3 之結果。藉由調整代表句擷取數量，每次重複執行十次取平均，在平均擷取數量下比較其代表句與人工標記答案之分數結果。從圖中可以觀察到方法 2 之分數隨著平均代表句擷取數量增加而降低，但逐漸趨於收斂。方法 3 則在平均擷取數量較低時，與方法 2 些許的不同，隨著擷取數量上升，卻同樣收斂於接近方法 1 的分數。由此可知方法 3 以義項類別之群集中心作為整體分群起始點時，其中心點並無法有效的代表各義項之語境，因為與分群中心較近之代表句所獲得的分數反而較低。方法 2 則符合離中心點越近之代表句，較能表達該群集義項的。

表四、「亞馬遜」代表句擷取分數評估（方法 1）

Macro-average	Macro-average 標準差	Weighted-average	Weighted-average 標準差
0.840	0.004	0.851	0.036
Accuracy	Accuracy 標準差	平均代表句數	平均代表句數標準差
0.856	0.031	483 句	35.364



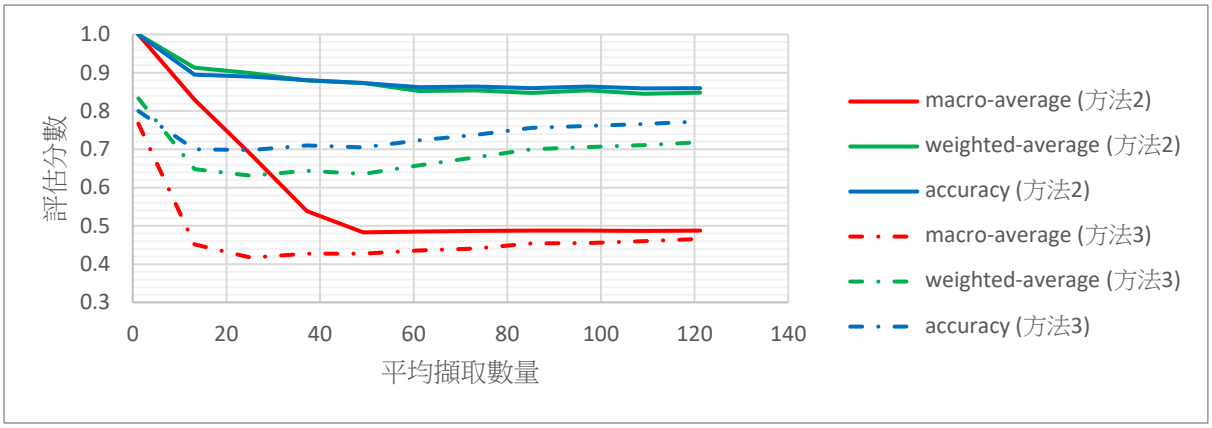
圖九、「亞馬遜」代表句擷取分數評估

接著是以「出發」作為代表，在最佳 purity 參數下的擷取結果，以相同方式進行比較。表五是以代表句擷取方法 1 的實驗結果。圖十則是以實線與虛線分別作為方法 2 及方法 3 之實驗結果。從中可以觀察到方法 2 與方法 3 在代表句擷取數量較低時，其分數皆高於擷取數量較高之分數，亦即距離群集中心越近之代表句越能代表該群集義項之語

境。但是兩者並無在擷取大量代表句時逐漸收斂至接近的分數，且 macro-average 與 weighted-average 相差約 0.236，推測是因為語料數據分布情形差異所造成之影響。

表五、「出發」代表句擷取分數評估（方法 1）

Macro-average	Macro-average 標準差	Weighted-average	Weighted-average 標準差
0.507	0.118	0.743	0.035
Accuracy	Accuracy 標準差	平均代表句數	平均代表句數標準差
0.807	0.007	3571.4 句	0.800



圖十、「出發」代表句擷取分數評估

（四）、代表句擷取之綜合比較

表六展示出各目標詞彙於最佳分群模型時，透過擷取方法 1 所得到的分數進行比較。從中可以發現同形異義之目標詞彙其指標分數皆高於一詞多義之目標詞彙，且三種指標分數皆相去不遠。反之，一詞多義之目標詞彙其 macro-average 與 weighted-average 大多存在一定落差，且 accuracy 分數皆低於同形異義之詞彙。若與表七一同觀察，可以發現多數屬於一詞多義之詞彙其例句分布對於三種指標的影響較同形異義之詞彙來的大。由此可推論若例句分布不均，難以將各義項有關聯，即一詞多義，之詞彙例句區分開來。

表六、各目標詞彙代表句擷取分數比較

目標詞彙	Macro-average	Weighted-average	Accuracy	擷取數量	擷取比例
亞馬遜	0.840	0.851	0.856	483.0	57%
蘋果	0.871	0.900	0.896	4456.0	100%
小米	0.954	0.961	0.961	300.6	40%
出入	0.722	0.777	0.763	935.6	90%
出發	0.507	0.743	0.807	3571.4	99%
壓力	0.394	0.514	0.651	6699.2	99%
東西	0.476	0.763	0.813	4199.8	70%

表七、各目標詞彙之義項比例分布

目標詞彙	義項 1	義項 2	義項 3	總句數
亞馬遜	雨林 51.6%	電商 48.4%		847
蘋果	科技產品 60.0%	水果 21.1%	報紙 18.9%	4367
小米	科技產品 78.4%	農作物 21.6%		667
出入	出外與入內 81.4%	不一致 18.6%		1037
出發	實際離開 81.6%	從某方面著手 18.4%		3574
壓力	緊張不安的狀態 64.9%	單位面積上所受之力 35.1%		6734
東西	物品 53.3%	位置 44.2%	東方與西方 2.5%	5933

四、結論

本研究目的是透過分群機制對於多義詞進行消除歧義。在實驗中能觀察到屬於同形異義之目標詞彙能有效地將多義詞不同的義項類別區隔開來，同時擷取出符合義項之代表句供使用者閱讀。然而屬於一詞多義之目標詞彙則無法得到好的成效，探究其原因，或許是因為一詞多義之義項是具有關連性的，因此難以僅透過現有方式在相關句語境上判斷出明顯的區別。

此外，多義詞的各個義項之例句在語料庫中分布不均，在機器學習上也會是個很大的影響，因此並非給定任意多義詞皆能得到良好的分群結果。如實驗結果表六與表七所示，屬於不同類別之詞彙，對於例句分布的情形存在不一樣的結果。

最後，依據實驗統計，我們仍可以發現特定分群模型在多數的詞彙下有較好的表現，對於本實驗以外的多義詞，可以優先推薦使用者以該分群模型進行分析。不過若要更精確地了解該分群模型是否適合使用者給定的詞彙，以目前實驗方式需要由使用者提供參考句，若能免去人工標記這繁雜的工作，或許未來能有更廣泛的應用。

參考文獻

[1] Bruce K. Britton, “Lexical ambiguity of words used in English text,” *Behavior Research Methods & Instrumentation*, 10(1), 1-7, Jan. 1978.

[2] John Lyons, *Semantics*, Cambridge: Cambridge University Press, Oct. 1977.

[3] 教育部，〈教育部重編國語辭典修訂本〉，網址：http://dict.revised.moe.edu.tw/cbdic/search.htm。

[4] 唐鳳，〈萌典〉，網址：https://www.moedict.tw。

[5] 中央研究院，〈中央研究院現代漢語標記語料庫〉，網址：http://asbc.iis.sinica.edu.tw。

[6] 吳美嫻，《〈長阿含經〉雙音詞研究》，碩士論文，國立東華大學中國語文學系，2010。

[7] 林香薇，〈閩南語歌仔冊中的多義詞「落 loh8」〉，師大學報：語言與文學類，第 61:2 期，頁 1-28，臺灣：國立臺灣師範大學，2016 年 9 月。

[8] 蔡宛玲，《漢語多義詞「跑」之結構及語意分析》，碩士論文，國立政治大學語言學研究所，2017。

[9] 許尤芬，《中文多義詞「發」之語義探討：以語料庫為本》，碩士論文，臺北市立教育大學華語文教學碩士學位學程，2012。

[10] Roberto Navigli, “Word sense disambiguation: a Survey,” *ACM Computing Surveys*, 41(2), 10, Feb. 2009.

[11] Michael Lesk, “Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone,” *Proceedings of the 5th Annual Conference on Systems Documentation*, Toronto, Ontario, Canada, 24-26, 1986.

[12] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to information retrieval*, Cambridge: Cambridge University Press, 356-360, 2008.

[13] 中文維基百科，〈維基百科:資料庫下載〉，網址：https://dumps.wikimedia.org/zhwiki。

[14] 教育部，〈新聞立場檢索技術獎金賽〉，
網址： https://aidea-web.tw/topic/b6abbf14-2d60-456c-8cbe-34fdfcd58967。

[15] Giuseppe Attardi, “WikiExtractor,” [Online]. Available: https://github.com/attardi/wikiextractor.。

[16] Carbo Kuo, “OpenCC,” [Online]. Available: https://github.com/BYVoid/OpenCC。

[17] 中央研究院詞庫小組，〈CKIP 中文斷詞系統〉，網址：http://ckipsvr.iis.sinica.edu.tw。

[18] ldkrsi, “jieba-zh_TW,” [Online]. Available: https://github.com/ldkrsi/jieba-zh_TW。

[19] Quoc Le, and Tomas Mikolov, “Distributed representations of sentences and documents,” *Proceedings of the 31st International Conference on International Conference on Machine Learning*, 32, 1188-1196, 2014.

[20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, “Efficient estimation of word representations in vector space,” *Proceedings of the International Conference on Learning Representations*, Scottsdale, Arizona, United States, 2013.

[21] James MacQueen, “Some methods for classification and analysis of multivariate observations,” *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297, 1967.

[22] Leonard Kaufman, and Peter J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, New Jersey: John Wiley & Sons, Inc., 1990.

[23] Wilm E. Donath, and Alan J. Hoffman, “Lower bounds for the partitioning of graphs,” *IBM Journal of Research and Development*, 17(5), 420-425, Sept. 1973.

[24] Miroslav Fiedler, “Algebraic connectivity of graphs,” *Czechoslovak mathematical journal*, 23(2), 298–305, 1973.

[25] Jianbo Shi, and Jitendra Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888-908, Aug. 2000.

[26] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss, “On spectral clustering: analysis and an algorithm” *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 849-856, Dec. 2001.

[27] Tian Zhang, Raghu Ramakrishnan, and Miron Livny, “BIRCH: an efficient data clustering method for very large databases,” *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, 25(2), 103-114, Jun. 1996.

[28] David Arthur, and Sergei Vassilvitskii, “K-means++: the advantages of careful seeding,” *Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms*,” 1027-1035, Jan. 2007.

[29] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to information retrieval*, Cambridge: Cambridge University Press, 279-285, 2008.