The 2014 Conference on Computational Linguistics and Speech Processing
ROCLING 2014, pp. 58-77

中文轉客文文轉音系統中的客語斷詞處理之研究

Research on Hakka Word Segmentation Processes in Chinese-to-Hakka Text-to-Speech System

黃豐隆[1]、余明興[2]、林昕緯[2]、林義証[3]

[1]國立聯合大學 資訊工程所，flhuang@nuu.edu.tw

[2]國立中興大學 資訊科學與工程所，msyu@dragon.nchu.edu.tw; lin@sinwei.tw

[3]建國科技大學 資訊管理學系，yclin@ctu.edu.tw

摘要

語言(Language)是文化傳承與推廣的首要工具，尤其是少數族群的語言，如：台灣的客語或原住民語言。臺灣的客家族群約佔總人口七分之一，為閩南語語系外之第二大族群。根據近年來相關臺灣客語使用狀況調查報告指出，阻礙客語傳承之主因是：不太會講。由於台灣學習環境使然，導致連客籍家庭的學童亦少能以客語說話、交談，具有聽、說客語能力者逐年下 降，能說客語的人口大量減少，台灣出現客語失聲、客家文化失傳之危機。

我們為了建置線上客語的數位學習系統，已開發出以大量合成單元為基礎的客語四縣腔及海陸腔的中文轉客文的文轉音系統(Hakka Text-to-Speech, HTTS)，以及相關的應用系統，如：線上國客雙語有聲詞典 [13]、國客雙語有聲地圖社群系統 [14]…等。

我們的系統，主要是提供不太會講客語或不會講客語的使用者來使用、學習客語。因此系統的輸入為「中文文句」，輸出為「客語語音」。這樣的操作設計，學習者或使用者能不需額外再學習客語輸入法、客語拼音，只需使用最熟悉的中文，即可透過本系統來學習客語。

為了更進一步改善與提升文轉音的效果，本論著重在改善系統中的客語文句分析模組的客語斷詞處理。在系統中，使用者輸入中文文句後，透過我們提出的客語斷詞方法，能將「中文文句」轉換為「客語文句及斷詞和詞性標記結果」。透過這個提升後的斷詞與詞性標記結果，來得到更佳的文句分析結果、提升文轉音中的文意正確性，如：韻律階層的求取、停頓類型的求取及讀音的求取。

本論文提出混合型的 N-Gram 序列分數算法，搭配中文斷詞模組及動態規劃演算法的客語斷詞方法。在嚴重資料稀疏的客語語料下，對中文轉客語斷詞結果的精確率有 80.78%。相較於傳統中文詞直翻客語詞的方法，已提升不少。

Abstract

Language is a major tool for cultural inheritance especially for the minority nationality, for example Hakka and aborigine language in Taiwan. As second ethnic besides Minnan dialect, the population of Hakka in Taiwan is one seventh. According to the recently reports of Hakka usage survey in Taiwan, the difficulties to inherit the culture of Hakka is missed in spoken Hakka language, the reason is the environments for learning and has led to the results of descending population for communicating by Hakka. It will become crucial for the cultural inheritance of Hakka.

Therefore, we has developed the Text-to-Speech method and system for Hakka language, and our goal is building environments for leaning the Hakka language, our some applied system such as:

"Web Hakka Phonetic Dictionary" [13] and "Blogging System of Bilingual Language by Integrating Mobile Cells and Google Map" [14] ,etc.

Our system is provided for users who interested in Hakka language, who can input the Chinese texts and system will output the speech of Hakka, users need not to learn the typing and phonetic writing of Hakka, and can take the advantage to learning Hakka with familiar language.

For the advanced improvements of Hakka Text-to-Speech, this article will emphasis on the word segmentation processing of Hakka text. In our system, when user enter the Chinese text, our proposed methods can convert the Chinese text to Hakka text and assign the part-of-speech for each Hakka text segments. By the better performance of text segments and part-of-speech in Hakka, We can improvements the Hakka text analysis module.

We proposed an hybrid N-gram sequence score, and Chinese word segmentation module developed by the dynamic programming algorithm, in the data-sparseness of Hakka corpus, the accuracy of Chinese to Hakka word segmentation is 80.78%.

Keywords: Hakka Text-to-Speech, Hakka Word Segmentation, Dynamic Programming, Hakka Text Analysis.

1. 緒論

一個斷詞系統的效果，通常跟語料的大小有關。但目前客語語料的收集非常困難，現有的電子資料，如：客委會初級、中高級的認證教材、教育部編著的國小客語教材…等，對於自然語言處理來說，資料規模仍然屬極小量語料。因此，想要建置出更多的客語語料，幾乎都需要從客語書籍、文章中，透過人工輸入、建置成電子檔的方式來取得。但有了這些文本資料只是第一步，後續仍有許多的處理工作，如：斷詞、詞性標記的處理，擷取出這些語言特徵後，才能做更進一步的分析與應用。

客語詞的判定是一件嚴謹的事情，理論上我們必須遵照詞的定義[1]來標記，但有極少數的情況下，我們仍會將詞組標記成一個客語詞，如：滑溜溜，在中文斷詞被斷為：滑/溜溜，我們視它是一個詞。而對於非客語語言專家的標記人員來說，其最有效率的方法，是透過具有平行資訊[2]的語料，先將中文語料輸入至中文的文句處理系統，取得中文的斷詞、詞性標記的特徵後，再對其對應的客語文章，以人工方式去判斷客語詞的邊界與詞性的標記。這個方法普遍被使用於同類型[3]的平行語料標工作記上，如 Tsai 的碩士論文 [1]也是用此方法。因為中文文句處理系統中，在文句斷詞資訊標記的技術方面已經相當成熟，而客語與中文的文法結構也相近，實際上中文的斷詞、詞性特徵，幾乎都能直接對應於客語詞。

目前客語語料的收集與建置，在學界有許多學者專家已積極的在做努力，建置出客語研究的相關基礎資料庫。如屏東教育大學的「學術研究基礎建置暨客家文化研究計畫 [2]」，他們歷時了至少三年的時間，在收集、建置客語語料及詞頻庫。這項創舉能有助於客語文句處理的發展，如：客語斷詞系統、客語文句分析系統、客語文句剖析系統、客語語音合成系統、智慧型的客語輸入法…等，都非常需要足夠的客語語料來支持其發展。

---

[1] 詞(Word)，是指最小、有完整明確意義且可以自由使用的中文語言單位。
[2] 具有中文文句和客文文句 1 對 1 對應的平行資訊的語料。
[3] 客語與中文都屬於漢語，文法結構幾乎相同，僅有少數俚語、特殊的客語構詞不同。

本研究旨在提出一個中文轉客文斷詞資訊的方法，針對嚴重資料稀疏的情況下，提出搭配中文斷詞模組與客語語言模型的混合式 N-Gram 序列分數的算法。透過兩階段方式，將中文文句以中文斷詞模組得到第一階段的斷詞及詞性標記結果後，再以國客語對照辭典找出所有可能被轉換的客語詞序列，以少量客語 Bi-gram、Uni-gram 語言模型為基底，搭配混合式 N-gram 序列分數的算法，找出分數最高的客語詞轉換序列，來得到第二階段轉換後的結果。

因礙於人力有限、語料收集的困難，仍有許多無法突破之處，如：語料的規模、人工標記資料的正確性。但使用本論文方法的客語斷詞法，以內部測試結果可知，若在訓練語料充足的情況下，能得到一個不錯的斷詞效能，其內部測試的精確度達 94.46%。相信在未來持續增加客語語料的規模後，本論文所提出的方法，效能會有更顯著的提升。

2. 文獻探討

2.1 中文斷詞

中文斷詞法每年都有新的研究與技術，甚至每年都會有舉辦斷詞比賽，知名的比賽如：SIGHAN 所舉辦的國際中文分詞競賽。第一屆競賽起始於 2003 年在日本札幌舉行。而之後每年都有相當多高手共襄盛舉。

在這麼多琳瑯滿目的斷詞技術中，常見的中文斷詞技術可分為三大類：(一)統計式斷詞 法、(二)法則式斷詞法和、(三)混合式斷詞法。

(A) 統計式斷詞法

統計式斷詞法藉由收集詞彙資料，如詞彙長度和詞彙出現的頻率或次數等統計上的資訊。然後系統運用此訓練資料經由演算法分析來取得斷詞序列。常見的演算法如 Xue 使用的 Maximum Entropy [15]，最後實驗得到最好的 F 分數是 94.98%，或是 Lo 使用的 Conditional Random Field [4]，最後實驗得到最好的 F 分數是 96.40%。這些演算法都是利用字元之間的資訊當作特徵。然後把斷詞問題轉換成為字元之間的分類問題。

而過去常被使用的演算法是 Hidden Markov Model。如 Fu 和 Luck [5]從訓練語料中統計詞頻、字元在詞中出現位置的次數等資訊，組合過後做實驗。得到 F 分數最好可達 93.7%。而 Lin 和 Chang [4]使用兩階段特製化的方式，藉著擴充觀測符號及狀態符號來改善隱藏式馬可夫模型的斷詞效果。得到 F 分數最好可達 96.3%。

(B) 法則式斷詞法

法則式斷詞法主要是根據一些經驗法則做為斷詞的標準，藉以達到較好的斷詞序列。常見的法則如「長詞優於短詞」、「與左邊詞的結合優於與右邊詞的結合」。而這類型的斷詞法常被參考的是 Chen 和 Liu [17]，他們在該篇論文中提出了六條法則(heuristic rules)，並且根據這些法則解決了歧義性的問題及剔除一些較不可能的詞彙組合，以完成中文斷詞的工作。在實驗上效果相當不錯。

不過，此種斷詞法容易受到詞典的好壞而影響效能。若是句子中出現未知的新詞彙時，則正確率就可能下滑。

(C) 混合式斷詞法

每種方式的斷詞法都有好壞、優缺點，因此後來學者們才會嘗試去混合兩種斷詞方式。早期的 Nie 等人[18]提出結合詞典、經驗法則及統計資訊來對中文斷詞。而令人印象深刻的是 Wu 和 Jiang [19]提出結合剖析器和斷詞器的方法，在斷詞時先使用查詞典來產生所有可能的斷詞組

合，再使用經驗法則剔除不可能的詞彙組合，然後再利用剖析器解決剩下的歧義問題，最後得到斷詞序列。在實驗上 F 分數高達 99%。

而比較不同的做法是 Gao 等人 [20]提出的專有名詞法則，這些法則幫助系統抓取未知的專有名詞如:人名、地名、組織名、外來語音譯人名。在最後實驗上 F 分數約為 96%左右。

2.2 客語斷詞

因客語語料稀疏的緣故，目前針對中文轉客文的相關研究非常少，客語斷詞系統的實做，先以(一)輸入為客語、(二)輸入為中文，分為兩大種類。第一類是直接針對客語文句做斷詞，第二類是針對中文文句翻譯成客語斷詞結果。

而目前仍沒有任何一篇是針對客語斷詞做深入研究的論文，其中對客語斷詞有做效能評估的論文，也僅有 Tsai 的碩士論文 [1]—基於隱藏式馬可夫模型之客語文句轉語音系統。顯見目前客語斷詞的研究，不管是語料的建置，還是斷詞的方法，仍非常多待探討與解決的問題。

(A) 輸入為客語

這一類的系統，適合具備客語輸入能力及熟悉客語的使用者，對於一般不熟悉客語的使用者而言，較不方便。這類系統常見的做法，是直接使用中文斷詞系統，對客文做斷詞。當然，這樣會有一些客語造字或客語用詞無法辨別的問題，針對這部份，是使用國客語對照辭典，來解決客語未知詞(Out of Vocabulary, OOV)問題。

如 Tsai 的論文 [1]，他們透過 Conditional Random Field 方法實做中文斷詞系統，並加入國客語對照外部辭典，配合客語構詞規則，實做出客語斷詞模組。最後的實驗效能，客語斷詞的 F 分數為 82.87%，客語詞性標記的 F 分數為 77.14%。

(B) 輸入為中文

這一類的系統，使用者不需使用客語輸入法，也不需熟悉客語，很適合客語初學者使用。這類系統常見的做法，是使用中文斷詞系統，先將輸入的中文文句斷詞，找出詞與詞性後，再將詞透過國客語的平行對照辭典，翻譯成客語詞。如本實驗室的線上客語語音合成系統，Wu [5]、Lo [6]的斷詞方法皆相同，都是使用 Jiang [7]所提出的中文斷詞系統，將中文文句斷詞後，再透過國客語對照辭典，將中文詞翻譯成客語詞。經測試後，其不含詞性標記的客語斷詞效能的 F 分數分別為 69.82%及 66.72%。

另一種是僅透過國客語對照辭典，將中文文句直翻成客語。如 Lee [8]，他們建置出一套國客語對照辭典，將輸入的中文文句字串切割成 1 到 4 字詞，並查找對照辭典、翻譯成客語詞。而他們沒有針對中文翻客語詞做效能評估，因此無法得知效果如何。

3. 準備工具及語料

3.1 中文斷詞工具

本論文的中文斷詞系統，是使用 Lai 於 2011 提出的「應用多詞及多詞性語言模型的中文斷詞及詞性標記方法 [9]」。此斷詞方法採用兩階斷式，第一階斷是斷詞，第二階斷是詞性標記。其斷詞的 F 分數有 96.69%，詞性標記的 F 分數 92.04%。

3.2 國客語對照辭典

我們所建置的國客語對照辭典，主要來源有：(一)客委會初級、中級暨中高級認證語料 [10, 11]、(二)台北市客委會-現代客語詞彙彙編。對於斷詞系統而言，辭典是決定正確率的重要因 素，理論上辭典越大，斷詞效能也就越好。因此除了現有的辭典來源外，我們也利用標記客語

文句斷詞答案的同時，找出一些尚未被收錄在辭典中的國客語對照之詞目。最後，也針對每個詞目，進行人工校正工作，去除重複或不合理的詞目以及標記拼音。

表一、國客語對照辭典資料樣貌

欄位	內容	說明
ID	13267	資料庫中 ID
Chinese	年輕人	中文詞用詞
Hakka	後生人	客語詞用詞
Pinyin	heu1 sang2 ngin3	客語拼音
Pos	Na	客語詞性
Pos_pattern	Na	中文詞性組
Hakka_pos_feq	25	客語含詞性詞頻
Hakka_nonpos_feq	25	客語不含詞性詞頻
Chinese_pos_feq	21138	中文含詞性詞頻
Chinese_nonpos_feq	21138	中文不含詞性詞頻

表二、2014 版，興大國客語對照辭典分佈統計及比較

字詞	2014 興大客語辭典	Wu[5]	Lo[6]	交大客語辭典[1]
一字詞	3457	780	838	6747
二字詞	20800	16362	16540	18078
三字詞	7309	5654	5826	5095
四字詞	4093	3769	3861	4217
五字詞	312	273	283	250
六字詞	84	80	84	80
七字詞	65	63	68	60
八字詞	9	12	14	14
總計	**36129**	**26993**	**27514**	**34541**

3.3 客語四縣腔語言模型的建置

語言模型是斷詞系統中，用來選擇斷詞結果的重要元件。而語料經過統計詞頻後，就能得到該份語料的機率分佈模型，即是語言模型。因此，語料越大，能得到的統計資訊越多，語言模型能包含的情況越多，效能也會越好。

但是，現今電腦上的客語語料仍非常匱乏，能用來建置語言模型的語料非常有限。因此，我們開發了一個半自動式的客語語料建置工具來建置客語語料，透過這個工具，可以將具平行資訊[4]的客語語料，標記出客語斷詞資訊，再透過該語料統計出客語 Uni-gram 及 Bi-gram 語言模 型。

[4] 具有中文文句和客文文句 1 對 1 對應的平行資訊的語料。

3.3.1 語料來源

我們用來建置客語語言模型的客語語料，主要來源是客委會四縣腔初級、中高級客語認證教材 [10, 11]。這份語料，句子數分別有初級 1678 句、中高級 4962 句，共 6640 句，每一句都有中文、客語的詞目、拼音及例句，如表：

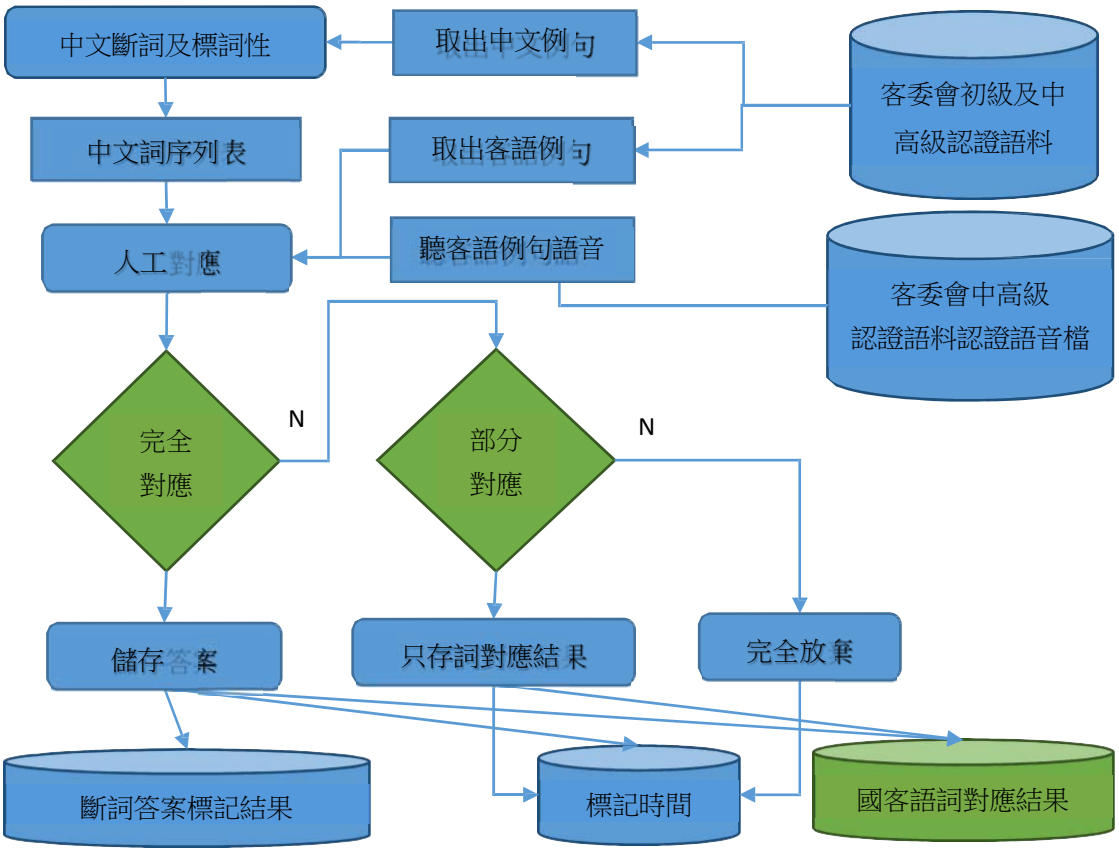
表三、客委會語料的資料樣貌

客語詞	光線
中文詞	光線
客語拼音	gong ` sien ˇ
中文例句	這個房間的光線不好，不適合做書房。
客語例句	這隻房間个光線毋好，毋適合做書房。

3.3.2客語斷詞標記工具及人工標記原則

客語與中文的文法結構相近，因此中文斷詞和詞性的標記，大部分都能與客語完全對應，僅有少部分的客語俚語或特殊用詞例外。而中文語料的處理，因目前中文斷詞系統的發展已相當成熟，因為中文語料的龐大，以規則法配合機率模型的混合式斷詞法所發展出來的中文斷詞系統，斷詞效能及詞性標記的 F 分數已達到 96.69%及 92.04%。因此都能直接以中文斷詞系統來得到可靠的斷詞及詞性特徵標記的結果。但客語文章的處理，因目前市面上及學界的客語斷詞系統仍處於發展中的階段，還沒辦法依賴任何客語斷詞系統來自動處理。因此我們開發出一套工具，以半自動的方法，快速的針對客語句子，做客語斷詞的標記。

3.3.2.1 客語斷詞標記工具介紹與操作



圖一、客語斷詞答案標記工具架構圖

我們將客委會四縣腔認證教材的客語語料 6640 句句子，依照句子的編號順序，由小到大分為語料 A 及語料 B，語料 A 有 4196 句，語料 B 有 2444 句，並再將 B 語料依編號順位分為 4 份，B1-B4，每份 611 句。

將 A、B 語料，分別找(標記者 1)一位碩士生、(標記者 2~5)四位大學在學生，使用本論文開發的客語斷詞答案標記工具，以半自動人工判斷方式，標記出客語斷詞的答案。其中每人的標記速度如表四：

表四、客語斷詞標記工具的標記時間統計，時間單位：秒。

標記者	1	2	3	4	5	總平均
語料編號	A	B1	B2	B3	B4	
處理句數[5]	4500	615	622	611	646	
儲存句數	4018	346	451	504	419	
總時間[6]	91798	28698	25297	18864	17219	
平均每句	20.39	46.66	40.67	30.87	26.65	33.04

表中可看出，標記速度平均 33.04 秒能完成一句。而標記完成的資料，會儲存其 1.中文句子、2.中文斷詞及詞性標記結果、3.客語句子、4.客語斷詞及詞性標記結果、5.客委會語料中的句子編號，等五個欄位資料，儲存結果如下表：

表六、客語斷詞標記結果的資料樣貌

中文句子	這個房間的光線不好，不適合做書房。
中文斷詞結果	這(Nep) 個(Nf) 房間(Nc) 的(DE) 光線(Na) 不(D) 好(VH) ，(COMMACATEGORY) 不(D) 適合(VH) 做(VC) 書房(Nc) 。(PERIODCATEGORY)
客語句子	這隻房間个光線毋好，毋適合做書房。
客語斷詞結果	這(Nep) 隻(Nf) 房間(Nc) 个(DE) 光線(Na) 毋(D) 好(VH) ，(COMMACATEGORY) 毋(D) 適合(VH) 做(VC) 書房(Nc) 。(PERIODCATEGORY)
句子編號	01-001

將這些標記完成的資料再經過一次經人工篩選後，最後確認有效句數為：A 語料 4018 句，B1-B4 語料共 1282 句，我們使用語料的分佈如表所示：

表七、客語語料的使用分佈

	訓練	測試
句數	4018	1282
詞數	45304	17646
字數	65572	25478

[5] 處理句數的統計，包含重新處理曾經放棄的句子，因此實際處理可能會比分配到的筆數多。
[6] 時間單位為秒。

圖二、標記工具操作畫面

3.3.2.2 標記原則

因為這份客語認證教材語料，專家們編審的主要目的是為了客語教學用，並不是為了要建立「國語/客語斷詞對應」的語料，所以在撰寫例句時並不會特別求強或注意到國客語的完全對應。

因此，我們在進行人工標記時也發現，其實大部分出現無法對應的情況，都可以以人工修改中文句子或用詞的方式，做適當的修飾與調整，來達到不影響文意、又能與客語句子對應完全的目的。但某些句子仍無法確認如何標記時，標記者也可選擇放棄該句的標記，或只存能對應到的詞，而不將這些未完成對應的句子視為斷詞答案。以下為標記時的 6 大標記原則：

原則 1：標記時，都以不修改客語句子為原則，但可跳過不影響文意的字串。

表八、標記原則 1 範例

中文	這泉水的水質很甜很清澈。
客語	這**窟**泉水个水質當甜當清。
中文改	這泉水的水質很甜很清澈。
客語改	這泉水个水質當甜當清。

此例子中，窟這個字只是指一窟井或一窟水池的意思，省略後也不至於影響整句文意。

表九、範例 1 標記後樣貌

中文改	這(Nep) 泉水(Na) 的(DE) 水質(Na) 很(Dfa) 甜(VH) 很(Dfa) 清澈(VH) 。(P)
客語改	這(Nep) 泉水(Na) 个(DE) 水質(Na) 當(Dfa) 甜(VH) 當(Dfa) 清(VH) 。(P)

原則 2：如果有很明顯且離譜的斷詞錯誤，要人工介入修正。

表十、標記原則 2 範例

原斷詞	這(Nep) 群(Nf) 小孩子(Na) ，(COMMACATEGORY) 每(Nes) **天都(Na)** 在(P) 沙洲(Na) 上(Nes) 玩(VC) 摔跤(VA) 。(P)
人工修正	這(Nep) 群(Nf) 小孩子(Na) ，(COMMACATEGORY) 每(Nes) **天(Nf) 都(Da)** 在(P) 沙洲(Na) 上(Nes) 玩(VC) 摔跤(VA) 。(P)

此例子中，「天都」一詞，被誤斷成一個普通名詞，這跟「天、都」意思不同，「天都」指的是地方名，而其正確斷詞應該斷成「天(Nf) 都(Da)」。

原則 3：可微調中文句子用詞及詞的順序以求對應到客語，但修改後的文意不能改變。

表十一、標記原則 3 範例 1

中文	今天的天氣很好，太陽下山以後就**可以看得到**滿天的星星。
客語	今晡日个天時當好，日頭落山以後就**看得著**滿天个星仔。
中文改	今天的天氣很好，太陽下山以後就**看得到**滿天的星星。
客語改	今晡日个天時當好，日頭落山以後就**看得著**滿天个星仔。

此例子中，若直接省略「可以」這個詞，也不會影響到原句的文意。

表十二、標記原則 3 範例 2

中文	古時候的人會觀察**天上的星宿變化**，來判斷人間的吉凶。
客語	上早个人會觀察**天頂星宿个變化**，來判斷人間个吉凶。
中文改	古時候的人會觀察**天上星宿的變化**，來判斷人間的吉凶。
客語改	上早个人會觀察**天頂星宿个變化**，來判斷人間个吉凶。

此例子中，客語「天頂星宿个變化」與中文「天上的星宿變化」雖然詞的詞順序不同，但中文句子中的「的」調換後，也不影響文意。而修改句子的動作，我們只建議修改中文，客文若非必要儘量保持原句。原因是較能保持客語句子原來的特性、結構、用語…等資訊。

原則 4：以客語句子為主體，發現中文詞和客語詞的對應有爭議時或太過模糊時，要找過一個最佳選擇的詞替換。

表十三、標記原則 4 範例

中文	古早的人若是看到流星雨，會想到**一些**壞兆頭。
客語	上早个人係看著星仔瀉屎，會想著**麼个**壞兆頭。
中文改	古早的人若是看到流星雨，會想到**什麼**壞兆頭。
客語改	上早个人係看著星仔瀉屎，會想著**麼个**壞兆頭。

在此例子中，客語的「麼个」翻成中文「一些」，依客語文意來看過於模糊。因此，可透標記工具中的詞典查詢功能，找到客語詞「麼个」能翻成的國語詞有哪些，發現到「什麼」這個中文詞最貼近文意。因此，手動將中文句子的「一些」改為「什麼」，重新與「麼个」配對。

原則 5：若沒辦法標記完一整句，但部分詞能對應，則選擇「放棄，儲存詞配對結果」。

表十四、標記原則 5 範例

中文	一身乾彆簡單粗陋的衣服都沒有能力買。
客語	一身腊食皮無才調買。

此例子中，我們發現這句除了「一/一、身/身、無/沒有、才調/能力、買/買」這些詞能對應外，其他詞皆無法非對應。因此此次的不納入正確答案範本內，但有部分詞能對應，也將這些詞的國客語對應結果儲存起來，待後續的工作中，仍可應用。

原則 6：配對時以「詞」為單位，不要以片語為單位。 通常，非成語的片語，用法可能只是偶然出現，因此儲存這類資料沒意義。

表十五、標記原則 6 範例

中文	現在外面既颳風又下雨，要怎麼樣回家？
客語	這下外背風合雨，愛仰般形轉屋下？

此例子中，發現到「既颳風又下雨」和「風合雨」看似能對應，但實際上可能只有在這句曾出現這種情況。因此，這種非成語的片語，也許只是偶然出現，我們不儲存此類的答案和詞配對。但依照原則 5，我們發現「要怎麼樣回家/愛仰般形轉屋下」可配對，成「要/愛、怎麼樣/仰般形、回家/轉屋下」。因此這句可儲存其「詞配對」成功的部份，但放棄儲存為斷詞答案。

4. 研究方法

本系統實驗有包括兩大種基底，(一)中文斷詞邊界優先、(二)客語詞邊界優先，經實驗發現第一種方法的外部測試正確率較高。但這可能跟標記語料時採用的：先中文斷詞、再人工對應客語詞的方法有關。但文章篇幅有限，本篇論文寫的實驗數據都以第一種方法為基底。

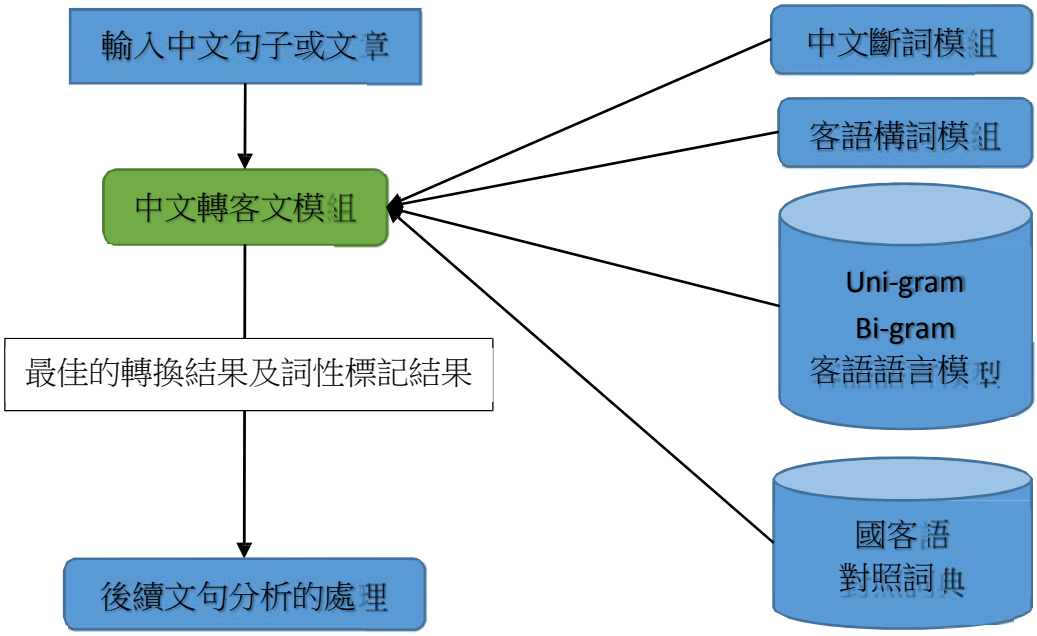
所謂(一)中文斷詞邊界優先，是先將輸入中文文句做斷詞，得到確定的詞邊界後，再以這些中文詞去查找國客語對照辭典，找出可能的客語詞轉換候選。而(二)客語詞優先，是先從國客語對照辭典中，找出所有可能被轉換的客語詞，最後再與中文斷詞邊界的詞一起做為候選詞。

兩種方法的不同之處：方法(一)的詞邊界已被固定，相對能找出的候選客語詞也較少、較侷限，方法(二)的詞邊界是自由的，可從國客語對照辭典及原中文斷詞斷的中文詞中，找出一條最佳的斷詞路徑。兩種方法其實各有好處，第一種方法在外部測試有較佳的正確率，第二種方法在內部測試有較佳的正確率。其意義是：如果在訓練語料充足的情況下，第二種方法會更 好。而第一種方法適合用在訓練語料稀疏的情況。

4.1 系統架構

在本論文的客語斷詞系統中，輸入是「中文文句」，輸出是「客語文句」的斷詞及詞性標記結果。簡而言之，以中文文句輸入後，經此客語斷詞模組處理，產生具有斷詞與詞性的客語文句的輸出。標記斷詞及詞性兩種特徵後的結果，可再使用文法剖析器分析、得到文法結構樹，做更進一步的文句分析處理。如：用於 Hakka Text to Speech (HTTS)中的停頓預估模型中，預測出句子中的 no break、minor break 及 major break 三種停頓類型，讓合出的語音可辨度更高、讓使用者能更輕易的聽懂句子內容。

因此，一個良好的客語斷詞系統是客語語言處理所不可或缺的，應是提升客語語音合成效果的重要因素。本系統中客語斷詞的架構，參見圖三。



圖三、客語斷詞模組架構圖

4.2 修改中文斷詞辭典

前文有提到，本論文採用第一種方法「中文斷詞邊界優先」為基底，因此為了讓一些客語用詞有機會成為被轉換的候選詞，我們將國客語對照辭典中，所有中文詞欄位的詞資料，都加入到中文斷詞辭典裡。如此做法能提高原本完全不會被找到的客語詞，有機會成為被轉換的候選詞。以下是一個例子：

表十六、中文斷詞邊界限制，範例 1

中文句子	泥鰍滑溜溜，
中文斷詞	泥鰍/滑/溜溜/，
中文翻客語	鰗鰍仔/滑/溜溜/，
正確答案	鰗鰍仔/滑溜溜仔/，

透過中文斷詞得到「泥鰍/滑/溜溜/，」的斷詞邊界，再透過以上方法找出最佳詞頻的客語詞並轉換為「鰗鰍仔/滑/溜溜/，」。但其實我們的國客語對照詞典中，有收錄「滑溜溜/滑溜溜

仔」這筆個國客語對照資料，但礙於中文斷詞邊界之限制，只能由一個詞「滑溜溜仔」轉成「滑/溜溜」兩個詞。此情況會直接的影響到轉換的正確率。

表十七、中文斷詞邊界限制，範例 2

中文句子	大家來去客家庄走一走。
中文斷詞	大家/來去/客家庄/**走/一/走/**。
中文翻客語	大家/來去/客家庄/**行/一/行/**。
正確答案	大家/來去/客家庄/**遶遶啊/**。

上面的例子「走一走」被斷成「走/一/走」，但實際上我們詞典有收錄「走一走/遶遶啊」這個對照詞組，但礙於中文斷詞邊界，我們沒辦法選到該詞組。

基於以上的原因，我們決定將國客語對照辭典中的中文詞欄位，加入到中文斷詞辭典。而這些新加入的中文詞，要決定出它們的詞頻大小，因為這兩個語料規模相差非常龐大，我們的中文斷詞辭典總詞頻數高達 462729801 個詞，而客語訓練語料僅有 47079 個詞。因此我們將中文斷詞辭典的平均詞頻取 Log 以 2 為底乘上 15(經實驗得到，15 有最佳的正確率)，再與國客語對照辭典的詞頻相乘，得到該詞新的詞頻。步驟如下：

1. 統計出中文斷詞辭典原始的分佈，我們得到其平均詞頻為 464 次。

表十八、中文斷詞辭典資料分佈

總詞數	995642
總詞頻	462729801
平均詞頻	464

2. 我們以下列公式，計算出每個要加入中文斷詞辭典中的中文詞，其新詞頻 $C(W_i)^*$，其中 $W_i \in (Word\ Length > 1)$：

$$C(W_i)^* = \log_2(464) * 15 * 〔C(W_i) + 1〕 \quad (1)$$

一個國客對照詞「植物/植物」，其詞頻轉換的例子：

表十九、國客語對照辭典的中文詞詞頻換算

中文	植物
客語	植物
未含詞性特徵的詞頻	3
新詞頻 $C(W_i)^*$	$Ceil(\log_2(464) * 15 * 〔3+1〕) = 532$

依照上述方法，產生以下新增候選列表：

表二十、客語詞新詞頻候選表

中文詞	詞頻
植物	532
雕刻	266
信仰	443
八仙	355
筵席	178
…	…

3. 將步驟 2 產生的新增候選列表的結果，加入中文斷詞辭典，若有相同的中文詞，則其詞頻相加。

最後我們評估修改前與修改後的中文斷詞系統，其正確率的差異。測試為外部測試，使用「中研院中文平衡語料庫 3.0」。如表二十一：

表二十一、中文斷詞辭典修改前後比較

	Precision	Recall	F-Measure
修改前	97.16%	96.21%	96.69%
修改後	97.15%	96.16%	96.65%

可看到 F-Measure 略低 0.04%，但改善了以下問題：

表二十二、加入客語詞後的中文斷詞辭典的改善

中文句子	泥鰍滑溜溜，
原中文斷詞	泥鰍/滑/溜溜/，
改善後中文斷詞	泥鰍/滑溜溜/，
原中文翻客語	鰗鰍仔/滑/溜溜/，
改善後中文翻客語	鰗鰍仔/滑溜溜仔/，
正確答案	鰗鰍仔/滑溜溜仔/，

原本的斷詞系統無法將「滑溜溜」判斷出來，修正辭典後已能正確斷出，並轉為客語詞「滑溜溜仔」。

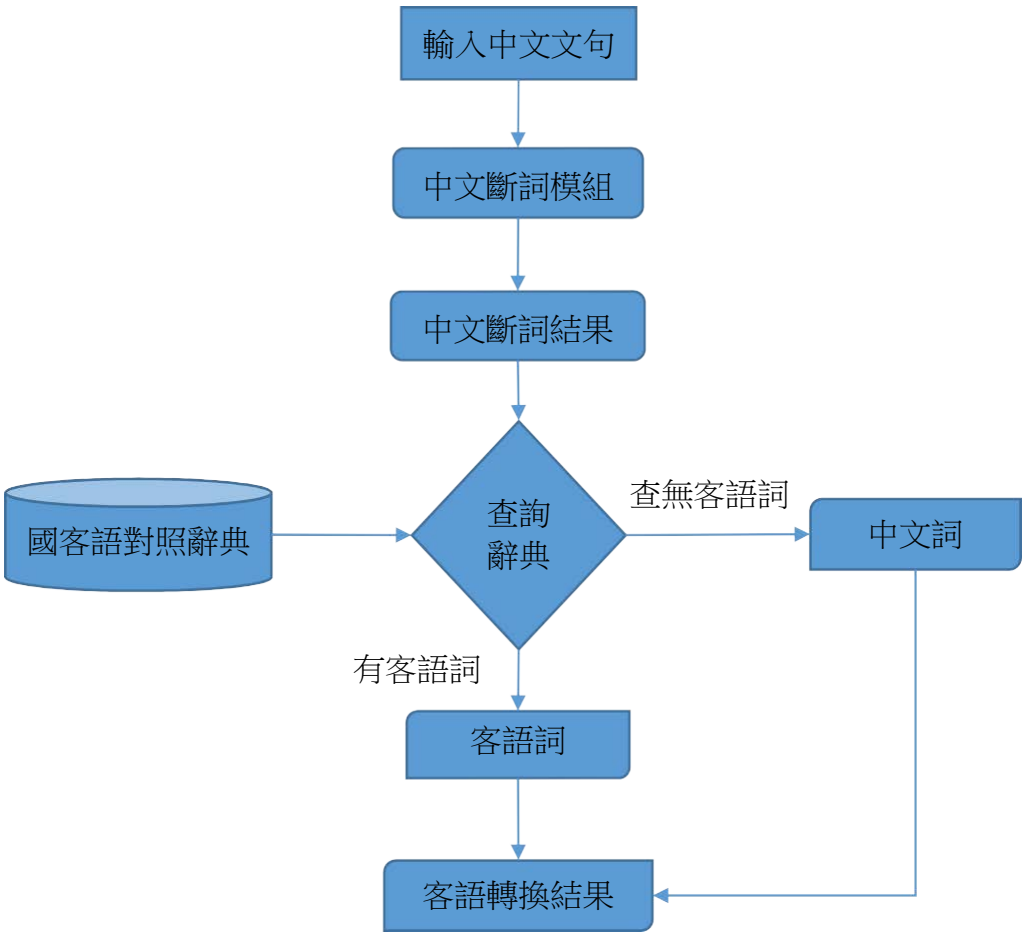
4.3 客語斷詞方法

4.3.1 中文斷詞搭配國客語對照辭典的直翻法

本方法的流程：

1. 透過中文斷詞系統得到中文斷詞及標詞性結果。
2. 查找國客語對照辭典，找出對應的客語詞，並選擇資料 ID 排序第一位者，並將中文詞轉換成該詞。
3. 若步驟 2 時查不到對應的客語詞，則沿用中文斷詞的結果。

圖四為此斷詞法的流程：



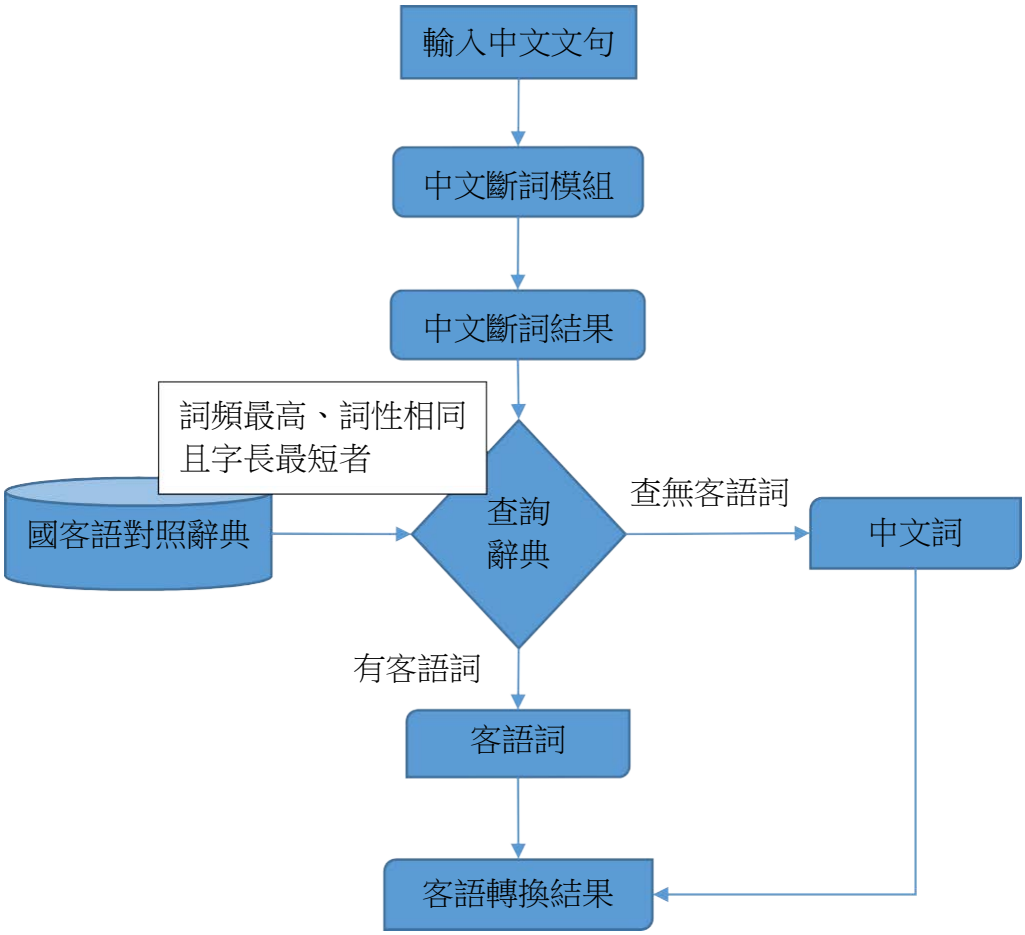
圖四、中文斷詞搭配國客語對照辭典的直翻法流程圖

4.3.2 中文斷詞搭配客語詞頻的直接翻譯法

以下是本方法的程式流程：

1. 透過中文斷詞系統得到斷詞及標詞性的結果。
2. 查找國客語對照詞典，找出對應的客語詞。
3. 挑選詞性與詞性標記結果相同者且詞頻最高、字長最短者為結果，若都找不到則以中文詞為結果。

圖五為本方法流程：



圖五、搭配客語詞頻的直接翻譯斷詞法

4.3.3 中文斷詞搭配客語 Uni-gram 及 Bi-gram 語言模型的混合式分數算法

本斷詞法，是先將輸入的中文句子，以中文斷詞模組做斷詞，得到一個確定的中文詞邊界後，再將這些中文詞查詢客語辭典，找出所有的可能候選詞。若找不到客語詞，則以僅有的中文詞當候選詞。最後將這些候選詞建立成一個所有斷詞路徑的有向圖，再以最短路徑演算法配合 Uni-gram 及 Bi-gram 混合式的分數算法，找出一條分數最佳的斷詞序列。

分數的計算方法：

本斷詞方法的分數計算方式，是混合式 Mix-gram(Uni-gram+Bi-gram)分數算法，如下列式子：

$$Seo\gamma e(<S>, W_1, W_2, W_3, \ldots, W_n) =$$

$$\alpha\gamma gmin 一〔 log_e〔 P(W_1|<S>) * P(W_1) 〕 + \sum^{n}_{i=2} Log_e〔 P(W_i|W_{i-1}) * P(W_i) 〕〕 \quad (2)$$

其中 $P(W_i|W_{i-1})$可利用 maximum likelihood estimation(MLE)來計算：

$$P(W_i|W_{i-1}) \text{ 起} \frac{C(W_{i-1},W_i)}{C(W_{i-1})} \quad (3)$$

若遇到訓練資料 $C(W_{i-1}, W_i) = 0$ 時，我們將 Bi-gram 機率以為 α 值代替，轉換如下：

$$P(W_i|W_{i-1}) = \begin{cases} \frac{C(W_{i-1},W_i)}{\sum_{W \in V} C(W_{i-1},W)}, & if\ C(W_{i-1}, W_i) > 0 \\ \alpha, & if\ C(W_{i-1}, W_i) = 0 \end{cases} \quad (4)$$

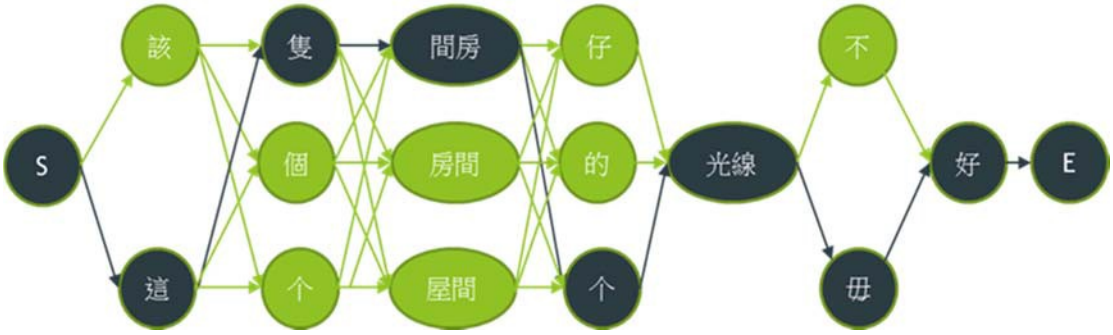
其中 $P(W_i)$：

$$P(W_i) = \frac{1+C(W_i)}{v+[\sum_{j=1}^{v} C(W_j)]} \quad (5)$$

$\sum_{j=1}^{v} C(W_j) = 47079$，為客語 Uni-gram 語言模型中的總詞頻數。

$V = 8931$，為 Uni-gram 語言模型的總 type 數(詞數)。

$\alpha = 10^{-3}$，為代替當 $C(W_{i-1}, W_i) = 0$ 時，用以替代 $\frac{C(W_{i-1},W_i)}{\sum_{W \in V} C(W_{i-1},W)}$的值，這是經由實驗得到，我們測試了 10^{-1}, 10^{-2}, 10^{-3}, …, 10^{-8} 等值。



圖六、中文斷詞搭配客語 Uni-gram 及 Bi-gram 語言模型的混合式分數算法，範例。

以上圖為例，如要計算上條斷詞序列 *Seoγe（S, 這, 隻, 間房, 个, 光線, 毋, 好）*，其 Mix-gram 的分數計算公式如下：

Seoγe（S, 這, 隻, 間房, 个, 光線, 毋, 好）=

[P（這|S） P（這）〕* [P（隻|這）* P（隻）〕* [P（間房|隻）* P（間房）〕* [P（个|間房）* P（个）〕* [P（光線|个）* P（光線）〕* [P（毋|光線）* P（毋）〕* [P（好|毋）* P（好）〕*

5. 系統效果評估

5.1 評估方法

本測試所輸出的句子，是中文未斷詞的句子，輸出客語斷詞及標詞性的結果。目前只測試單純客語分詞的效能。效能評估的計算方法，我們使用精確率(Precision)、召回率(Recall)、以及 F-分數(F-score)來評估系統的效能，這三種方法的定義如下所示：

$$\text{精確率} = \frac{\text{系統正確斷出的詞數}}{\text{系統斷出的總詞數}} \quad (6)$$

$$\text{召回率} = \frac{\text{系統正確斷出的詞數}}{\text{標準答案的總詞數}} \quad (7)$$

$$\text{F 分數} = \frac{2 * \text{精確率} * \text{召回率}}{\text{精確率} + \text{召回率}} \quad (8)$$

基於下列原因，除了上述的評估方法外，我們進一步運用編輯距離演算法(Levenshtein Distance)，評估轉換後的客語句子相似度(Similarity)。

原因 1：

中文翻客文的處理，是一個中文詞對多種可能客語字詞的問題，且客語常有一意多詞的問題，如：中文「收到」客語可翻成「收」或「收着」，若正確答案標記為「收着」，但系統輸出為「收」，在斷詞評估角度看，兩者詞不同，正確率為 0，但以相似度來看，兩者僅差一個字，相似度仍有 50%。

因此，除了斷詞邊界的評估標準外，其字串轉換後的相似度也可以是一個評估效能的方法。且客語文句的文法結構與中文幾近相同，沒有翻譯後文法結構對齊的問題，因此可用此方法計算其相似度。

原因 2：

本論文的斷詞系統，是用於客語語音合成系統中，因此，翻譯後的字串，距離標準答案的相似度越高，能將字念對的可能性也越高。

此演算法計算兩個字串 A、B 間，由字串 A 轉換成字串 B 的最小編輯距離(Insertions, Deletions 或 Substitutions)，計算方式如下：

$$D(i,j) = \min \begin{cases} D(i-1,j) + InsertCost(target_i) \\ D(i-1,j-1) + SubstituteCost(source_j, target_i) \\ D(i,j-1) + DeleteCost(source_j) \end{cases} \quad (9)$$

其中：

$$SubstituteCost = \begin{cases} 0 & if\ target[i] = source[j] \\ 1 & otherwise \end{cases}$$

$$InsertCost = 1$$

$$DeleteCost = 1$$

並由以下公式，將距離轉換成 0 到 1 之間的值，即為 A、B 字串間的相似度：

$$Similarity(A,B) = 1 - \frac{D(A,B)}{Max\ Length(A,B)} \quad (10)$$

5.2 測試語料

本實驗所使用的客語斷詞訓練與測試語料，是採用客委會四縣腔初級及中高級語料中的例句，共 6640 句句子。我們將這份語料，分為語料 A 及語料 B，語料 A 有 4196 句，語料 B 有 2444 句，並再將 B 語料依編號順位分為 4 份，B1-B4，每份 611 句。

而這些標記完成的資料，再經過一次經人工篩選後，確認有效句數為：訓練語料 4018

句，測試(B1-B4)語料共 1282 句，我們使用語料的分佈如表所示：

表二十三、客語語料的使用分佈

	訓練	測試
句數	4018	1282
詞數	45304	17646
字數	65572	25478

5.3 評估結果與討論

5.3.1 中文斷詞搭配國客語對照辭典的直翻法

實驗 A：使用 Lo[6]的國客語對照辭典。 **實**

驗 B：使用 Wu[5]的國客語對照辭典。

實驗 C：使用本論文所建置的國客語對照詞典。

表二十四、直接翻譯斷詞法效能評估-內部測試

	Precision	Recall	F- Measure	字串相似度
實驗 A	68.41%	69.12%	68.76%	73.29%
實驗 B	72.66%	73.42%	73.04%	75.68%
實驗 C	75.02%	75.80%	75.41%	78.36%

由此實驗可得知，詞典的校正與詞目的增加，能顯著的改善客語斷詞系統的效能。

5.3.2 中文斷詞搭配客語詞頻的直接翻譯法

實驗 A：輸入中文文句，評估其中文轉客語的斷詞效能。 **實**

驗 B：輸入中文文句，評估其中文轉客語及詞性標記效能。

表二十五、中文斷詞搭配客語詞頻的直接翻譯法-斷詞及詞性標記內外部測試結果

		Precision	Recall	F- Measure	字串相似度
實驗 A	**訓練**	88.16%	87.48%	87.82%	91.08%
	測試	80.32%	79.08%	79.69%	83.68%
實驗 B	**訓練**	87.46%	86.78%	87.12%	-
	測試	79.90%	78.66%	79.27%	-

以實驗結果來看，顯示客語詞頻的應用能顯著的提升選詞的正確率。但外部測試的正確率仍偏低，已接近未使用詞頻特徵的結果。此情況的原因，是因為用來統計客語詞頻的語料仍不足，因此造成統計資料稀疏的問題。但從實驗結果的特性觀察到，若能持續的增加客語語料、建置出更多的客語詞頻，能提升中文轉客文系統的效能。

而就目前的窘況而言，提升正確率的方法，僅能靠規則法(Rules-base)，如找出客語的構詞規則和文法規則，來提升客語斷詞的正確率。客語構詞部份，是下一階段即將進行的工作。

5.3.3 中文斷詞搭配客語 Uni-gram 及 Bi-gram 語言模型的混合式分數算法

實驗 A：輸入中文文句，評估其中文轉客語詞的斷詞效能。

實驗 B：輸入中文文句，評估其斷詞及詞性標記效能。

表二十六、中文斷詞搭配客語 Uni-gram 及 Bi-gram 語言模型的混合式分數算法

		Precision	Recall	F- Measure	字串相似度
實驗 A	**訓練**	94.46%	93.73%	94.10%	96.17%

	測試	80.78%	79.53%	80.15%	84.04%
實驗 B	**訓練**	93.96%	93.24%	93.60%	-
	測試	80.37%	79.11%	79.74%	-

加入 Bi-gram 後，內部測試的效能有顯著的提升，但外部測試僅略升 0.46%。原因是因為客語語言模型資料稀疏的關係，許多 Bi-gram pattern 未出現在客語 Bi-gram 語言模型中，或就算出現在 Bi-gram 語言模型中，也不符合句子實際的狀況，這是資料稀疏的問題。

而在下一階段的工作中，我們將要增加客語構詞規則以及持續增加客語語料，客語構詞包括了：重複、附加、附合、合併…等構詞規則。

6. 結論及未來工作

本論文針對中文轉客文文轉音系統(Hakka Text-to-Speech System, HTTS)中的客語斷詞處理，已提出一個基礎的研究架構。但因為目前客語電子語料有嚴重不足的問題，對於本論文所探討的主題而言，是一項非常艱困的挑戰。研究之初，我們並沒有客語斷詞語料可使用，僅有少量一句句未處理的國客語對照文句。因此，我們投入了大量時間在客語語料的標記、建置及辭典的校正、標音。我們也持續的從國小客語教材、客語朗讀比賽文章…等電子文本中，人工建置出更多的客語斷詞語料及客語新詞。而目前用來測試的語料，皆是我們自行建置的，所以我們也非常需要更多不同來源的客語斷詞語料，做更公證客觀的效能評估、比較。

本論文在客語斷詞方法方面，不同於過去的架構，我們提出了使用客語 Uni-gram 及 Bi-gram 語言模型的混合式斷詞序列分數算法，可看見在內部測試的評估上，顯示在語料充足的情況 下，將會有不錯的斷詞表現。但詞性標記的正確率僅能做為參考，因為我們自己標記產生的標 準答案，其詞性大部份都是依照中文斷詞系統給出的詞性為主，除離譜的錯誤外，很少當下進 行修正。因此，語料的斷詞、詞性標記資訊，仍需要經專家再一次做更嚴謹的校正。

本論文提出混合型的 N-Gram 序列分數算法，搭配中文斷詞模組及動態規劃演算法的客語斷詞方法。在嚴重資料稀疏的客語語料下，對中文轉客文的外部測試精確率有 80.78%，內部測試有 94.46%。相較於傳統中文詞直翻客語詞的方法，已獲得提升。相信未來持續增加客語語料的規模後，使用本論文所提出的方法，效能會有更顯著的提升。

客語斷詞的應用層面極廣，不僅止使用於我們中文轉客文文轉音系統中的文句分析模組，還可獨立用於客語的數位學習、客語文句處理、客語語音辨識…等領域。本論文提出的研究方法，應能提供未來客語斷詞相關研究做為基礎與參考。

我們接下來要進行的工作有：

1. 持續擴充國客語對照辭典。
2. 加入客語構詞規則。
3. 最佳化語言模型平滑化問題，如：Good-Turing Katz、Kneser-Ney。
4. 持續標記、建置客語語料。

致謝

客家委員會提供給本論文實驗用之客語認證語料以及獎助部份經費，特此致謝。

參考文獻

[1] 蔡依玲，基於隱藏式馬可夫模型之客語文句轉語音系統，國立交通大學電信程所碩士論 文，2009。

[2] 鍾屏蘭、江俊龍，學術研究基礎建置暨客家文化研究計畫，屏東教育大學客家文化所計畫成果報告書，2009。

[3] 羅永聖，結合多類型字典與條件隨機域之中文斷詞與詞性標記系統研究，國立台灣大學資訊工程所碩士論文，2008。

[4] 林千翔，*Chinese Word Segmentation using Specialized HMM*，國立中央大學資訊工程所碩士論文，2005。

[5] 吳俊毅，線上客語語音合成系統中產生韻律訊息之研究，國立中興大學資訊科學與工程所碩士論文，2010。

[6] 羅丞邑，以資料探勘之技術解決線上客語語音合成系統中多音字發音歧義之研究，國立中興大學資訊科學與工程所碩士論文，2011。

[7] 江昶毅，應用多種特徵的中文斷詞及詞性標記方法，國立中興大學資訊科學與工程所碩士論文，2010。

[8] 李雪貞，客語語音合成之初步研究”，國立臺灣科技大學資訊工程所碩士論文，2002。

[9] 賴亦傑，”應用多詞及多詞性語言模型的中文斷詞及詞性標記方法”，國立中興大學資訊科學與工程所碩士論文，2011。

[10] 客委會出版，客語能力認證基本詞彙-中級、中高級暨語料選粹四縣版上冊。

[11] 客委會出版，客語能力認證基本詞彙-中級、中高級暨語料選粹四縣版下冊。

[12] 林東毅，客語文句翻語音系統之實作，國立交通大學電信工程所碩士論文，2007。

[13] 黃豐隆，線上國客雙語有聲詞典建置之研究，全國計算機會議(NCS-2009)，台灣，2009。

[14] 黃豐隆，國客雙語有聲地圖社群系統，聯合大學資工所客委會計畫成果報告書，2013。

[15] Nianwen Xue, *Chinese Word Segmentation as Character Tagging*, Computational Linguistics 2003 February, Vol. 8, No. 1, pp.29-48.

[16] Guhong Fu and K.K. Luke., *A Two-Stage Statistical Word Segmentation System for Chinese*, Proceeding of The Second SIGHAN Workshop on Chinese Language Processing 2003, Vol. 17, pp.156-159.

[17] Keh-Jiann Chen and Shing-Huan Liu, *Word Identification For Mandarin Chinese Sentences*, Proceedings of COLING 1992, pp.101-107.

[18] Jian-Yun Nie, Marie-Louise Hannan, and Wanying Jin, *Unknown Word Detection and Segmentation of Chinese Using Statistical and Heuristic Knowledge*, Communications of COLIPS 1995, Vol. 5, pp.47-57.

[19] Andi Wu, Zixin Jiang, *Word Segmentation In Sentence Analysis*, International Conference on Chinese Information Processing in Beijing China 1998, pp.169-180.

[20] Jian-feng Gao, Mu Li, and Chang-Ning Huang, *Improved Source-Channel Models for Chinese Word Segmentation*, the 41st Annual Meeting on Association for Computational Linguistics 2003, Vol. 1, pp.272-279.