國立台北大學資訊工程學系專題報告
法律常用語辭典建構

專題組員:陳佑豪、吳承翰、張博崴、薛信瑒

專題編號:PRJ-CSIE-113-005
執行期間:2024 年 7 月 至 2025 年 5 月

摘要

一個高效的資料庫系統是發展進階智能服務的基礎。然而，由於法律用語斷詞具有（1）專用性、（2）組合性、（3）層次性等特性，傳統斷詞工具（如 Jieba）難以準確處理這類複雜詞彙，進而影響法律資訊檢索與文本分析的精確度，限制法律智能服務的進一步發展。為解決此問題，我們提出一套 JT-NG-IDF 自動化辭典建構方法，以協助斷詞工具提升法律用語斷詞的準確性。結果顯示 JT-NG-IDF 在長詞辨識與補充法律專業詞上效果最佳，能有效補足既有法律辭典之不足。

1. 簡介

隨著資料庫技術的成熟，許多領域都已建立相當好用的資料庫系統，並以此類系統為根基，提供切合使用者期待的智能服務，如：交易資料庫系統與喜好商品推薦服務。然而，中文法律文書資料庫仍停留在基本的字串比對，使得法律文書進階分析與服務一直未能有很好的智能服務發展，其中一個關鍵就在於法律文書的斷詞困難。

正確的斷詞是文書索引、分析的基礎，然而，透過資料觀察，我們發現法律文書的用詞具備以下三種與一般文書用詞顯著不同的特性：（1）專用性：某些詞語在使用一般斷詞工具時不易被正確斷出，但在法律語境中具有特定意義，因此需要被正確辨識，例如：「受領」、「催告」、「孳息」等；（2）組合性：由兩個或以上常見詞組合而成，雖然各成分在一般語境中可單獨出現，但在法律語境中卻具有特定整體意涵，應作為一個詞處理，例如：「相當因果關係」、「損害賠償」；（3）層次性：詞語本身呈現出語義層級的結構，從基本概念延伸出更細緻的法律適用情境，並形成不同的法律意涵與適用條件，如：「損害賠償」、「損害賠償責任」、「連帶損害賠償責任」。由於上述這些特性，使得現行的自然語言斷詞系統（例如 Jieba[1]），雖然在一般文本的斷詞很有幫助，但在處理法律專業字詞時常會遇到斷詞錯誤，導致語意理解錯誤，進而影響斷詞系統在法律資訊檢索、判決書分析及法律文件處理的精確度。

為了更好地處理法律用語斷詞問題，利用法律辭典改善斷詞是很直接又有成效的作法，例如將《裁判書用語辭典》讀進 Jieba 斷詞。然而，現有的法律辭典也缺少許多法律用詞，例如司法院裁判書用語查詢系統中並未收錄如「連帶損害賠償責任」、「情事變更」、「履行輔助人」及「無權處分」等詞彙。

因此我們希望藉由建立一個法律常用語辭典以幫助提升現有斷詞方法在法律專業領域的準確性；同時，不同於研究[2]使用人工篩選法律常用字詞，本專題也希望建立電腦自動化篩選方法，以讓本方法能夠在將來法律專業領域受到世代更迭後還能有效率地自動化篩選出法律專業字詞。我們預計開發一套基於文件分析建構辭典的自動化方法， 透過結合 Jieba[1]、TF（Term Frequency）、IDF（Inverse Document Frequency）[3]、N-gram [4]及資料分析等技術，從大量法律文本中提取專業詞彙，建構高品質的法律專業用詞辭典，並利用這一辭典來提升斷詞品質。

2. 相關研究

研究[5]使用 Jieba 斷詞搭配 TF-IDF 技術[3]進行專有詞篩選，在斷詞的結果交由人工判定篩選出重要詞，並篩選出 TF-IDF 高的詞製作出字典，再把篩選出的字典讀進 Jieba[1]以改進下次的斷詞成果，最後反覆迭代多次。這樣的處理方式依然高度仰賴人工篩選，且對於人工專業仍具備高度要求，因此，我們希望讓本方法能夠在將來法律專業領域受到世代更迭後還能有效率地自動化篩選出法律字詞。

3. 研究方法

我們的方法旨在從眾多法律判決書中提取法律常用詞彙，核心概念是透過多層次分析，先對文本進行初步分詞，並根據各詞出現的頻率（TF）篩選出判決書常用詞，接著分析詞與詞之間的共現頻率，將應屬於同一專業詞的片段合併生成更符合法律語義的長詞（N-gram），如「損害賠償」，最後檢查詞的普遍性（IDF）篩選出具法律意義而非案件意義的法律術語。

其中，在判決文書資料集裡，我們關注的詞是屬於 TF 高且 IDF 低的詞，因為這類詞彙在單篇文書中頻繁出現，且在整體語料中也很常見，通常代表具有穩定語義的法律常用詞。例如「合意解除」、「附隨義務」，這些詞不僅出現頻率高，也具有通用性，是我們最主要保留的對象。

因此，我們以 Jieba－tw、TF、IDF、N-Gram 的概念為基礎，建構一個新的方法 JT-NG-IDF，來擷取判決文書中的重要詞彙，此 JT-NG-IDF 方法包含三大步驟，分別處理基本斷詞與資料清理、長詞生成、代表詞擷取。

階段 1： 基本斷詞與資料清理(Jieba-tw、Term Frequency，JT)

在這個階段，我們以常用的文書處理技術處理台灣法院的判決書文本，將判決書文本使用 Jieba-tw [2]進行分詞，並進一步去除停用詞、單字詞、數字、符號及英文字。此外，在實務上，判決書中往往包含大量的序數前綴、時間前綴接數字再接單位量詞（如“第 1 條”、“民國 108 年”等），此類用詞經由 jieba-tw 分詞將產生大量破碎且意義相對模糊的分詞，造成後續分析困擾，因此，我們以格式規則方法將其排除。

此後，基於重要詞彙具備一定程度重複出現的特性，我們根據詞頻（TF）做過濾，擷取前 k（%）高頻詞作為關鍵詞，以便後續分析可以著重在重要詞彙上。

其中，詞頻的計算我們採用為每一篇文書計算一詞彙出現的次數，篩選出每篇文件中的高頻詞，直接將其視為關鍵字。此方法的好處是著重在個別判決書的詞彙出現頻率，則傾向於選出較為多樣且具變化性的詞彙，使得結果詞彙中既有辭典外的法律詞彙出現比例相對提升。

$$K_1 = \bigcup_d \{t | rank_d(t) \le k \cdot |t \in d|\},$$

$rank_d(t)$為文件d中的詞按照$TF(d,\cdot)$由大到小排序後，詞彙t的名次（最高頻 = 1）

$$TF(d,t) = \frac{n(d,t)}{\sum_i n(d,i)},$$

其中$n(d,t)$是詞彙 t 在文件 d 中出現的次數。

階段 2： 長詞生成(N-gram，NG)

在此階段，我們目標為從大量法律文本中自動擷取潛在的長詞組合，藉此補足現有法律專業辭典在長詞上的不足。實務上觀察可見，法律語言具有高度的組合性與層次性，例如「相當因果關係」、「信賴保護原則」等語塊詞，常由兩個以上常見詞語組成，卻攜帶不可分割的法律語義。為了擷取此類詞語，我們採用 N-gram 合併關鍵詞集 D 中的破碎短詞：我們從 Bi-gram 起步，進一步推展至 Tri-gram、Four-gram，逐步擴充詞組長度，期望可以涵蓋如「交通事

故調查書」、「附隨義務存在與否」等長詞組，並視實際應用需求彈性調整至更高階層，如 Five-gram 以上。

需要注意的是，在詞組合併中，由於不同的字後面可接續的詞的數量變化極大，舉例來說，一些詞如「合意」後續選擇較少（如「合意解除」），而其他短詞如「違反」可能有大量可能後接詞（如「違反法律」、「違反合約」、「違反規定」等），為了控制詞組合併的準確度與穩定性，我們設計動態門檻調整詞彙併條件，使得詞彙多寡與語義可信度的平衡可被兼顧。動態門檻的核心是以平均（1/M）為基礎，其中 M 代表當前詞後面可接續的候選詞數量。這種設計的邏輯在於，當候選詞較少時（如 M=2），語義連結通常較強，我們希望保留較高相關性的組合；而當候選詞眾多時（如 M=20 ）這些候選詞之間的語義連結較弱，應採取較低的門檻以減少錯誤過濾，如：

假設觀察詞”信賴”的候選延伸詞和其條件機率有：

- 保護：0.50
- 關係：0.30
- 基礎：0.20

因為候選數 M=3，根據動態門檻計算公式，門檻為 1/3＝0.33，基於此門檻，「信賴保護」會合併。

然而這種門檻也帶來潛在的問題：隨著 M 的增加，1/M 的門檻會趨近於 0，如M=50 時，門檻僅為 0.02，這將導致許多語義弱的組合被錯誤地合併，為解決這個問題，我們引入了一個固定門檻（如 0.1）作為下限，這種方式確保了在 M 極大時，門檻也不會過低。

因為法律詞彙本身常具有語義上的層次性，在此基礎上，我們設計了遞進式的詞組合併規則，以對應此種語義層次，具體而言，Tri-gram 合併僅針對已完成 Bi-gram 合併的詞對進行，亦即當前兩詞已被判定為語意緊密並成功合併後，才會進一步檢查是否能納入第三個詞構成更長詞組。Four-gram 的合併亦是建立在 Tri-gram 基礎上，逐層推進。這種由下而上的漸進式方式有助於在資料中穩定取得語義連貫的長詞組合。

我們定義一個合併條件機率門檻 θ_n（對 n-gram 設定的門檻），並用 K_2 表示透過 N-gram 擴展後的詞集。則：

- 合併條件：

對於任意詞組 $w_1, w_2, \ldots, w_n \in K_1$，若：

$$P(w_n \mid w_1,\ldots,w_{n-1}) = \frac{Count(w_1,\ldots,w_n)}{Count(w_1,\ldots,w_{n-1})} \geq \theta_n$$

則定義該詞組為合法合併詞，並加入擴展集 K_2。

- 階段二生成的詞集合：

$$K2 = K_1 \cup \{ w_1 \cdots w_n \mid \forall\, n \geq 2,$$

$$P(w_n \mid w_1,\ldots,w_{n-1}) \geq \theta_n,$$

$$w_1,\ldots,w_n \in K_1 \}$$

其中合併順序遵守階層性：只有 $w_1w_2 \in K_2$ 時才會考慮合併 $w_1w_2w_3$。

階段 3：代表詞擷取

在此階段，我們確認被擷取的關鍵詞是否具有法律及情境的通用性。具體而言，根據觀察，我們發現判決書中存在一些詞彙是偏向案件特性而非法律意義的用詞，如“西瓜刀”，這類的詞彙可能在極少數文書中以極高的頻率出現，且不屬於停用詞，因此被認定為關鍵詞。然而，這類詞彙並不具高度的法律及情境普遍性，因此，並不適合被納入法律用語辭典。為了過濾掉這種詞彙，我們評估每個關鍵詞的 IDF，僅將 IDF 低於門檻 λ 的詞推薦納入法律用語辭典 $Dict$：

$$Dict = Dict \cup \{t \mid t \in K_2, log(\frac{N}{A(t)}) \leq \lambda\}$$

4. 實驗

4.1 實驗設定

本研究以「交通事故民事賠償」為核心概念擴充，自司法院資料開放平臺提取 2022 年的資料為研究案例，具體篩選條件為：凡案由包含「損害賠償」、「侵權損害賠償」、「過失傷害」、「過失致重傷」、「公共危險」或內文包含「車禍」、「交通事故」、「損害賠償」或「慰撫金」之非憲法、行政、刑法或懲戒類型的判決，為本次專題所採用之資料，共 201,197 筆。

本研究設計 JT、JT-IDF、JT-NG-IDF 三種方法在各自表現最佳的參數條件下進行分析。具體設定如下：

1. JT 方法：取每篇文書中前 0.3% 的高頻詞彙。
2. JT-IDF 方法：取每篇文書中前 11% 的高頻詞彙，並擷取出 IDF 值小於 4.9 的詞彙。
3. JT-NG-IDF 方法:取每篇文書中前 11% 的高頻詞彙，經過 N-gram 合併後，擷取出 IDF 值小於 4.9 的詞彙。

為了評估並量化每個階段產出的字典之成效，我們從兩個面向採用六種指標進行評估：（1）與答案辭典進行比對：我們使用之答案辭典包括 tidy_append、law_terms_index，其中 law_terms_index 為司法院裁判書用語辭典資料庫中的法律用詞，而 tidy_append 是較為生活化之用詞，並用辭典與我們建構的辭典進行比較，評估召回率（Recall，R）、精確率（Precision，P）與 F 值（F1 Score）；（2）對料想外詞語進行檢視：邀請法律專業人士對建構辭典中未被答案辭典收錄的詞語進行評估，判斷其是否屬於法律專業用語，若是則記為增補率（Supplementation Rate，SR），若為一般詞或法律文書普遍用詞(但不具備特定法律意義)則記為冗餘率（Redundancy Rate，RR），若為錯誤詞則記為錯誤率（Error Rate，ER）。

tidy_append			
	JT	JT-IDF	JT-NG-IDF
P	27.93%	78.87%	69.88%
R	33.36%	80.86%	81.55%
F1	30.40%	79.85%	75.27%
SR	*0.97 % (30 字)	10.98% (85 字)	12.91% (159 字)
RR	*3.66 % (113 字)	62.79% (486 字)	51.54% (635 字)
ER	*0.39 % (12 字)	26.23% (203 字)	35.55% (438 字)

表一、自動化辭典建構方法在短詞法律辭典上的表現

law_terms_index			
	JT	JT-IDF	JT-NG-IDF
P	2.86%	7.02%	7.45%
R	4.10%	8.65%	10.53%
F1	3.37%	7.75%	8.73%
SR	*0.26% (11 字)	*1.29% (44 字)	*1.36% (53 字)
RR	*28.93% (113 詞)	*89.69% (3055 字)	*79.94% (3108 字)
ER	*0.29% (12 字)	*4.79% (163 字)	*5.04% (196 字)

表二、自動化辭典建構方法在長詞法律專業辭典上的表現

（一）$P(\alpha,\beta)=\frac{|\alpha\cap\beta|}{|\alpha|}$

（二）$R(\alpha,\beta)=\frac{|\alpha\cap\beta|}{|\beta|}$

（三）$F1(\alpha,\beta)=2\times\frac{P(\alpha,\beta)\times R(\alpha,\beta)}{P(\alpha,\beta)+R(\alpha,\beta)}$

（四）$SR(\alpha,\beta)=\frac{|LT(\alpha-\beta)|}{|\alpha\setminus\beta|}$

（五）$RR(\alpha,\beta)=\frac{|CTSLC(\alpha-\beta)|}{|\alpha\setminus\beta|}$

（六）$ER(\alpha,\beta)=\frac{|ET(\alpha-\beta)|}{|\alpha\setminus\beta|}$

其中α為建構字典之收錄詞集，β為答案辭典之收錄詞集，LT（Legal Terminology）篩選出法律專用語，CTSLC（Common Term Served as Legal Context）篩選實為一般詞但可作為法律內文使用之詞語，ET（Erroneous Term）篩出錯誤詞。

4.2 實驗結果與討論

表一與表二分別比較 JT、JT-IDF、JT-NG-IDF 三種方法對照 tidy_append、law_terms_index 兩種辭典的表現。結果顯示，在六項指標上，JT-IDF、JT-NG-IDF 表現都明顯優於 JT，這表示考量 IDF 可以篩選出更具法律意義而非案件意義的詞彙。比較 JT-IDF 與 JT-NG-IDF，JT-IDF 在短詞（tidy_append）上似乎優於 JT-NG-IDF（有較高的 P、R、F1 及較低的 ER），而在長詞（law_terms_index）上則表現不如 JT-NG-IDF（以絕對詞數觀察，特別是在表一中，SR 的差異），這驗證了我們引入 N-gram 模型的初衷，即為了強化對組合詞的擷取能力，能具擷取法律專業用語的能力。另外，比較表一與表二，由於 law_terms_index 為法律專業辭典，收錄較多法律專業長詞，使得三種方法在 P、R 及 F1 的表現都極差，但這更凸顯了 JT-IDF 及 JT-NG-IDF 在 SR、RR 的優秀表現，顯示現有法律辭典不夠完備的困境。

表三呈現了 Jeiba 搭配使用不同字典時，平均斷詞錯誤次數，其中，Raw 為 Jieba 常配搭的一般辭典。從其結果可以看出，當在 Raw 基礎上加入我們產出的字典（JT-NG-IDF）後，Jieba 的斷詞精確度有所提升，平均錯誤從 6.65 降至 6.57，變異數也從 18.3954 下降至 17.4538。另一方面，Raw 加入現有的法律專業字典（Tidy+Law）時，平均錯誤雖然僅略降，但變異數顯著降低，從 17.45 降至 6.72，顯示斷詞結果更為穩定；然而，再次加入我們產出的字典（JT-NG-IDF）時，因為 N-gram 合併機制會產生部分錯詞，導致平均錯誤其實是略為上升的，但由於涵蓋了更多法律專業用詞，可使整體變異數進一步降低至 5.1262，這說明斷詞結果趨於一致，呈現出更高的一致性與穩定性。

依據上述結果，我們認為在尚未具備專業辭典的情況下，加入我們所建構的字典整體上能有效提升斷詞精確度與穩定性；然而，當已有一定規模的專業辭典時，效果仍有限，可能仍須思考如何降低 N-GRAM 的錯誤率。

Jieba 斷詞	平均錯誤 （變異數）
Raw	6.6538 (18.3954)
Raw+JT-NG-IDF	6.5769 (17.4538)
Raw+Tidy+Law+JT-NG-IDF	5.3846 (5.1262)
Raw+Tidy+Law	5.1923 (6.7215)

表三、Jeiba 搭配使用不同字典時，平均斷詞錯誤次數

5. 結論

本專題觀察法律用語的特性，建立自動化辭典建構方法，能有效協助改善現有斷詞工具在法律斷詞上的限制；此自動化辭典建構方法亦可用於未來法律專業辭典的自動更新及維護，幫助建構、維護高效法律資料庫系統，以提供智能法律文本檢索及分析應用。

6. 銘謝

感謝指導教授於專題的製作過程中給的建議與幫助，引導我們進行方向的同時保留足夠多的空間讓我們嘗試各種可能的方法，除了定期的討論外也撥出了額外的時間協助專題的進行。也感謝實驗室的學長姊，於製作中遇到問題時提供了許多幫助。

參考文獻

[1] Sun, J. (2012). Jieba: Chinese text segmentation [Software]. Available from https://github.com/fxsjy/jieba

[2] APCLab. (2017). Jieba-tw: Traditional Chinese text segmentation [Software]. Available from https://github.com/APCLab/jieba-tw

[3] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, Vol. 28 No. 1, (pp. 11–21). https://doi.org/10.1108/eb026526

[4] Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[5] Xu, B., Hu, W., Fu, Q., & Zhang, Q. (2021). Research on Text Information Mining Technology of Substation Inspection Based on Improved Jieba. In *2021 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)* (pp. 379–383). IEEE. https://doi.org/10.1109/ICPICS52425.2021.9616554