台北大學資訊工程學系113年度專題

法律常用語辭典建構

陳佑豪、吳承翰、薛信瑒、張博崴

一、簡介

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，我們提出一套自動化辭典建構方法，以提升法律用語斷詞的準確性。

專用性　僅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。

組合性　由兩個或以上常見詞組合而成，並在法律語境中有特定整體意涵，例如：「相當因果關係」、「損害賠償」。

層次性　詞語本身呈現出語義層級的結構，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」。

二、研究方法

JT 階段：基本斷詞與資料清理 → **NG 階段**：長詞生成 → **IDF 階段**：代表詞擷取

為了對應法律的組合性及層次性，使用N-gram嘗試將多個詞彙合併

在資料清理後，使用TF高分篩選，確保詞在一定出現頻率以上，以選出常出現的詞彙

根據JT結果計算條件機率 ⇨ 根據機率分布控制合併門檻 ⇨ 設定門檻下限 ⇨ 遞進式合併

使用IDF低分門檻，為了對應法律詞彙在文章間的普遍性，使出現的字在一定量文章中出現

D={...,"信賴","保護","原則","",...}
計算基於 "信賴" 的條件機率

" 信賴 " 延伸：
保護：0.50
關係：0.30
基礎：0.20
門檻為 1/3＝0.33
" 信賴保護 " 會合併。

" 違反 " 延伸
有50個候選詞
→門檻僅為0.02

"信賴", "保護", "原則"→
"信賴保護", "原則"→
"信賴保護原則"

三、評估指標

我們採用六種指標進行評估：

α為建構辭典之收錄辭集
β為答案辭典之收錄辭集
LT篩選出法律專用語
ET篩選出錯誤詞
CTSLC篩選實為一般詞但可作為法律內文使用之詞語

精確率：$P(\alpha,\beta)=\frac{|\alpha\cap\beta|}{|\alpha|}$

召回率：$R(\alpha,\beta)=\frac{|\alpha\cap\beta|}{|\beta|}$

F值：$F1(\alpha,\beta)=2\times\frac{\mathrm{P}(\alpha,\beta)\times\mathrm{R}(\alpha,\beta)}{\mathrm{P}(\alpha,\beta)+\mathrm{R}(\alpha,\beta)}$

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

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

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

四、實驗結果

TidyAppend（Tidy）：主要由短詞組成的法律辭典
LawTermsIndex（Law）：主要由長詞組成的法律專業辭典
Raw：Jieba常用一般辭典

Tidy

	JT	JT-IDF	JT-NG-IDF
P	27.93%	78.87%	69.88%
R	33.36%	80.86%	81.55%
F1	30.4%	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)

Law

	JT	JT-IDF	JT-NG-IDF
P	2.86%	7.02%	7.45%
R	4.1%	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%(1202字)	*89.69%(3055字)	*79.94%(3108字)
ER	*0.29%(12字)	*4.79%(163字)	*5.04%(196字)

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