中文易混淆字集的建置與應用之研究

【文／語文教育及編譯研究中心助理研究員　白明弘】

一、錯別字是網路時代的特色之一

自有文字以來，錯別字即存在於書籍中。許多古書中的錯別字留傳了下來，至今成為通假字的來源之一。現代由於網路媒體發達，網路閱讀者也參與了訊息的產出，文字訊息常常快速埋沒在網路的洪流之中。訊息的流通性固然提高了，但關注的效期卻降低了。因此，訊息校正的重要性逐漸被忽略。根據國外研究，搜尋引擎中輸入的關鍵詞高達 26%包含錯別字，錯別字已成為網路時代的特色之一。

二、自然語言研究力挽錯別字狂瀾

錯別字除了不便於閱讀外，在資訊處理上也是一大麻煩。一篇包含錯別字的文章，可能因此無法被檢索到；而使用者輸入的檢索條件如果包含錯別字則找不到正確的文件。幸而自然語言研究的發展，逐漸填補了錯別字所造成的問題。現今大部分知名的搜尋引擎（包括 google、bing、度等），都已支援錯別字更正建議。而文件編輯器（包括 MS word 及 OpenOffice），也都支援錯別字與文法偵錯與更正建議的功能。以英文來說，錯別字偵測的正確率大約可達 99%。

錯別字的偵測與修改，基本上採取雜訊通道模型（Noisy Channel Model）。以英文錯字 fand 為例，想像某個英文詞在傳送時，其中一個字母傳送錯誤。但接收端無從知道原來是什麼詞。更正的方法只能一一取代其中字母去猜測。例如 fand 可能是 band, hand, land, sand, fend, find, fond, fund 或 fans 的誤寫。接著從這 9 個詞中挑出最可能的詞。例如 band 誤寫成 fand 的機率高不高？這個問題可以簡化成 b 被誤寫成 f 的機率高不高？這種字母 x 被誤寫成 y 的機率表叫做混淆矩陣（confusion matrix）。以英文來說，混淆矩陣大約要 26x26 筆資料表，這個表中記錄著每個字母被誤寫成另一個字母的機率。

三、中文錯別字問題仍懸而未決

儘管英文錯別字的問題已幾近於解決，但中文錯別字偵測連 80%的正確率都難達成。Hsieh 等提到原因之一是中文字集十分龐大。中文光是常用字就有 5,000 個，一個簡單的混淆矩陣就至少需要 2 千 5 百萬筆機率資料，而訓練這些機率資料則需要至少數億筆的錯別字語料。

四、中文易混淆字集

由於混淆矩陣難以實行，退而求其次是將矩陣簡化成對應字集。以「不同凡想」的「想」字為例，混淆矩陣將所有中文字都當成可能替代字，總共有 5,000 種替代情況。而易混淆字集則只考慮最可能的幾個替代字，例如：「想」 {餉響享饗像湘箱香項向相息…}。假設每一個字都只建立 20 個最可能的易混淆字，則 5,000 個中文字大約需要 10 萬筆機率資料。雖然數量還是很大，但已經在可處理的範圍了。

五、教學上的應用

中文易混淆字集除了應用在錯別字自動偵測之外，還可以利用中文易混淆字集從大量語料中自動抽出中文錯別字的實例。這些錯別字實例在教學上是非常實用的資料。依照語文教育文獻的建議，除可做為「集中識字法」的素材，幫助學生歸納和辨析字的用法之外，也可以發展一些中文文字辨析遊戲協助學生學習正確的用字。

【資料來源】

Hsieh, Y.-M., Bai, M.-H., Huang, S.-L., & Chen, K.-J. (2015). Correcting Chinese Spelling Errors with Word Lattice Decoding. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 14(4). https://doi.org/10.1145/2791389