基於隱藏式馬可夫模型的中文熟語自動糾錯新方法
A Novel Chinese-Idiom Automatic Error Correction Method Based on the Hidden Markov Model

張榕彬 **Rongbin Zhang**[1*], 桂安露 **Anlu Gui**[1], 曹鵬 **Peng Cao**[1]
吳凌峰 **Lingfeng Wu**[1], 黃鳳 **Feng Huang**[1], 李家暉 **Jiahui Li**[2]
[1]Department of Electronic Engineering, Shantou University, Shantou, China
[2]College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou, China
{rbzhang, 24algui, 22pcao, 22lfwu, 22fhuang1}@stu.edu.cn
qsx20240853@student.fjnu.edu.cn

摘要

在日常學習和使用中，受各種錯別字和光學字元辨識誤差等影響，中文熟語的拼寫經常出現錯誤。實現熟語自動糾錯是中文自然語言處理的重要任務之一，有助於提升中文文本品質和漢語學習效果。現有的編輯距離方法和自定義詞典方法，都存在糾錯能力受限、計算效率較低、靈活性不足等問題。鑒於此，本文提出一種基於隱藏式馬可夫模型（hidden Markov model，HMM）的中文熟語自動糾錯方法，將熟語錯誤的產生過程用 HMM 模型來建模，把熟語糾錯問題轉化爲錯誤熟語與合法熟語之間的匹配問題。通過構建合法熟語表和漢字混淆集，開發一個熟語自動糾錯原型系統，並完成性能測試。實驗表明，與現有方法相比，本方法模型簡單、參數少、計算複雜度低，具備更強的糾錯能力和參數健壯性，能夠更靈活地糾正多樣化類型的熟語錯誤，具有較高的應用潛在價值。

Abstract

Spelling errors in Chinese idioms frequently occur due to various types of misspellings and optical character recognition errors in daily learning and usage. Achieving automatic error correction for Chinese idioms is one of the important natural language processing tasks, as it helps improve the quality of Chinese texts as well as language learning. Existing methods, such as edit distance and custom dictionary approaches, suffer from limited error correction capability, low computational efficiency, and weak flexibility. To address these limitations, this paper proposes a novel automatic error correction method for Chinese idioms based on the hidden Markov model (HMM). Specifically, the generation process of idiom spelling errors is modeled using an HMM, transforming the idiom correction problem into a matching task between erroneous idioms and legitimate idioms. By constructing a legitimate idiom table and a Chinese character confusion set, a prototype system for idiom correction was developed, and performance testing was completed. Experiment results demonstrate that the proposed model is simpler with fewer parameters and has lower computational complexity while exhibiting stronger error correction capability and parameter robustness as compared to existing methods. It can more flexibly correct diverse types of idiom errors, showing high potential application value.

關鍵字：中文熟語、自動糾錯、隱藏式馬可夫模型

Keywords: Chinese idiom, Automatic error correction, Hidden Markov model

1 緒論

中文是世界上最古老的文字之一，其歷史悠久，沉澱了豐富的詞彙和熟語。其中，熟語（idiom）是指人們從日常生活經驗中總結出來的短語或短句，通常言簡意賅、含義深刻，例如成語、慣用語、歇後語、諺語、格言、詩詞名句等。然而，由於熟語結構特殊、用法靈活，在日常學習和使用時，容易出現拼寫錯誤。常見的熟語拼寫錯誤來源，有近形字、同音字、近音字、光學字元辨識錯誤、鍵盤誤操作，最終表現爲漢字重複、缺失和誤用等。研究實現中文熟語自動糾錯，是自然語言處理領域的問題之一，對於檢驗中文教學效果、提升中文文本品質等具有重要意義。

本文以隨機過程、時間序列和機率統計的視角，重新思考熟語拼寫錯誤的產生原因和應對方法，開發一種靈活糾正多樣化錯誤類型的中文熟語自動糾錯輕量型方案。我們的靈感來自於隱藏式馬可夫模型（hidden Markov model，HMM）在語音辨識等領域的應用 (Rabiner, 1989)。在針對孤立詞（isolated words）的語音辨識中，HMM 模型用於建模語音信號的序

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 124–132
November 20-22, 2025

		Transformation			
Error	Correction	Correct Letter	Error Letter	Position (Letter #)	Type
acress	actress	t	—	2	deletion
acress	cress	—	a	0	insertion
acress	caress	ca	ac	0	transposition
acress	access	c	r	2	substitution
acress	across	o	e	3	substitution
acress	acres	—	s	5	insertion
acress	acres	—	s	4	insertion

Figure 1: 與錯誤詞 acress 編輯距離爲 1 的合法候選詞示例。截圖自 (Jurafsky and Martin, 2024) 的 Appendix B

列特徵，利用模型的狀態轉移機率和觀測機率等參數，經運算後對說話者表達的詞語或句子進行辨識。Lin et al. (2012) 曾利用 HMM 模型實現一個英文單詞拼寫檢查器的簡單原型系統，其思想與 HMM 模型在語音辨識中的應用有異曲同工之妙。HMM 模型在時間序列分析中所展現的靈活性和健壯性，爲本方法提供了寶貴靈感和理論基礎。

現有的熟語拼寫自動糾錯方法，主要有兩種。第一種是基於編輯距離（edit distance，也稱 Damerau-Levenshtein 距離）(Hodge and Austin, 2003; Wang et al., 2014) 的方法，常用於英文單詞拼寫檢查 (Jurafsky and Martin, 2024; Norvig, 2016; Revathi et al., 2023)，隨後也用於俄語 (Varlamova et al., 2023)、孟加拉語 (Khairul Islam et al., 2019)、緬甸語 (Mon et al., 2021)、印度尼西亞語 (Soleh and Purwarianti, 2011)、印地語 (Jain and Jain, 2014) 等單詞糾錯。該方法需要根據給定的錯誤詞，產生具有編輯距離的可能詞作爲候選集，進而在合法詞彙表中進行查詢和匹配。其中，編輯距離是指字元插入（insertion）、字元刪除（deletion）、字元替換（substitution）、相鄰字元換位元（transposition）等基本操作的次數，如圖1所示。然而，由於優選集的大小隨著編輯距離的增加呈現指數級增長，該方法通常只能處理具有較小編輯距離的拼寫錯誤，其糾錯能力受到極大的限制。

第二種是自定義詞典方法，即首先在詞典中列出常見的錯誤詞及其對應的正確詞，進而將錯誤詞與詞典中的樣本逐一進行比對。典型案例是微軟公司 Word 辦公軟體自帶拼寫檢查器 (Microsoft-Corporation, 2025) 的成語糾錯功能，如圖2所示。然而，該方法只能處理已被收錄的特定拼寫錯誤，因而靈活性低，糾錯能力有限。例如，根據圖2的詞典，該方法能夠將「冰上巴蕾」正確地糾正爲「冰上芭蕾」，但無法對「冰上巴雷」進行檢錯和糾錯，因爲詞典並未收錄「冰上巴雷」這一錯誤詞。

針對現有方法在糾錯能力、計算效率和靈活性等方面的局限，我們使用 HMM 機率模型，

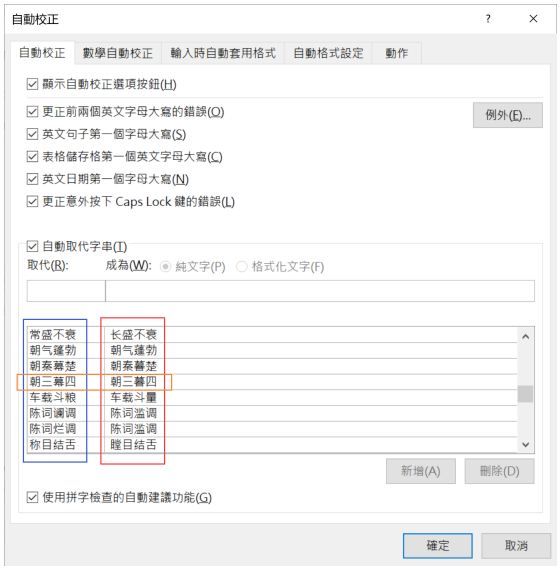

Figure 2: 微軟公司 Word 辦公軟體拼寫檢查器只能糾正已被收錄的錯誤成語。截圖自：Microsoft Word -> 檔案 -> 選項 -> 校訂 -> 自動校正選項 -> 自動校正

實現一種中文熟語自動糾錯輕量級方法，以期對於多樣化類型的熟語使用錯誤都能夠靈活地自動糾正。本論文的貢獻如下：

1) 以隨機過程、時間序列和機率統計等新視角，重新看待中文熟語糾錯問題，使用 HMM 模型進行問題建模和求解。

2) 在問題建模中，對於 HMM 模型的轉移機率和觀測機率等關鍵參數，都巧妙地賦予物理意義，使得所提出的自動糾錯演算法具有很強的合理性和可解釋性。

3) 與基於編輯距離的傳統方法相比，本方法的靈活性更強，可糾正不同類型的中文熟語使用錯誤，而不局限於簡單的少量字元替換、刪除或插入錯誤。

4) 與基於自定義詞典的傳統方法相比，本方法的糾錯能力更強，而不局限於詞典中預先收錄的特定錯誤詞。

5) 本文所提出演算法模型簡單、參數少、計算效率高、可擴展性強，可根據實際需求即時更新合法熟語表，可靈活地應用於成語、慣用語、歇後語、諺語、格言、詩詞名句等的自動糾錯任務。

2 提出方法

2.1 HMM 基礎

HMM 模型最初由 Leonard E. Baum 等提出，後來 Lawrence R. Rabiner 等進行深入研究 (Rabiner, 1989)。該模型是一種雙重隨機過程，表現在其狀態轉換過程是隱藏的、無法直接觀測的，而可觀測事件的隨機過程是隱藏

狀態轉換過程的隨機函數 (Zong, 2024)。

習慣上，使用五元組 $\lambda = (N, M, \mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ 來表示一個離散平穩 HMM 模型，參數包括：

1) 隱藏狀態的個數 N；

2) 觀測狀態的個數 M；

3) 轉移機率矩陣（transition matrix）$\mathbf{A}$。該矩陣是 $N \times N$ 方陣，其第 i 行第 j 列元素 $[\mathbf{A}]_{i,j}$ 表示任意時刻 t 由隱藏狀態 s_i 轉移爲下一時刻 $t+1$ 的隱藏狀態 s_j 的機率；

4) 觀測機率矩陣（也稱混淆機率矩陣、發射機率矩陣，observation/confusion/emission matrix）$\mathbf{B}$。該矩陣有 N 行 M 列，其第 j 行第 k 列元素 $[\mathbf{B}]_{j,k}$ 表示任意時刻 t 由隱藏狀態 s_j 產生觀測狀態 v_k 的機率；

5) 初始機率向量（initial probability vector）$\boldsymbol{\pi}$。該向量是長度爲 N 的行向量（row vector），其第 i 個元素 $[\boldsymbol{\pi}]_{1,i}$ 表示在初始時刻 $t=1$ 系統處於隱藏狀態 s_i 的機率。

HMM 模型涉及評估（evaluation）、解碼（decoding）、訓練（training）等三個基本問題 (Rabiner, 1989; Zong, 2024; Stamp, 2021)。其中，評估問題描述爲：給定 HMM 模型 $\lambda = (N, M, \mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ 和觀測狀態序列 $\mathbf{o} = o_1 o_2 \cdots o_T$，計算在該模型下產生序列 $\mathbf{o}$ 的機率 $\Pr\{\mathbf{o}|\lambda\}$。該問題可使用前向演算法（Forward Algorithm）進行求解。本文所提出的方法將使用評估問題來建模並求解。

2.2 熟語自動糾錯的 HMM 建模

本方法的切入點在於爲合法熟語表中的所有熟語分別建立各自的 HMM 模型，而關鍵在於如何合理設置每個 HMM 模型的參數。

我們將待糾錯的錯誤熟語 $\mathbf{w}$ 視爲觀測狀態序列，例如含有拼寫錯誤的「冰上巴雷」。假設所使用的合法熟語表共有 N_{idiom} 個熟語，例如 THUOCL 詞庫提供的合法成語表 (Han et al., 2016) 共有 $N_{\text{idiom}} = 8519$ 個成語。記第 n 個合法熟語爲 $\mathbf{c}_n\,(n = 1, 2, \cdots, N_{\text{idiom}})$，其對應的 HMM 模型爲

$$\lambda_n = (N_n, M_n, \mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n). \tag{1}$$

模型 λ_n 產生觀測狀態序列 $\mathbf{w}$ 的機率爲 $\Pr\{\mathbf{w}|\lambda_n\}$。此時，一種合理的策略是取

$$\mathbf{c}_{\text{opt}} = \arg\max_n \Pr\{\mathbf{w}|\lambda_n\} \tag{2}$$

作爲最佳糾錯建議。式 (2) 所使用的最優準則，可以看成是錯誤熟語與所有合法熟語之間的最佳匹配。[1]

[1]可根據實際需要使用其它最優準則。例如，若改用 $\mathbf{c}_{\text{opt}} = \arg\max_n \Pr\{\lambda_n|\mathbf{w}\}$，可根據 Bayes Formula，將其等效地表示爲 $\mathbf{c}_{\text{opt}} = \arg\max_n \Pr\{\mathbf{w}|\lambda_n\}\Pr\{\mathbf{c}_n\}$，此時只需在式 (2) 的基礎上，加入先驗機率知識 $\Pr\{\mathbf{c}_n\}\,(n = 1, 2, \cdots, N_{\text{idiom}})$，即每個合法熟語在漢語環境下被使用到的機率（可用語料中出現的頻次表示）。本文以式 (2) 的最優準則爲例，進行闡述和展示。

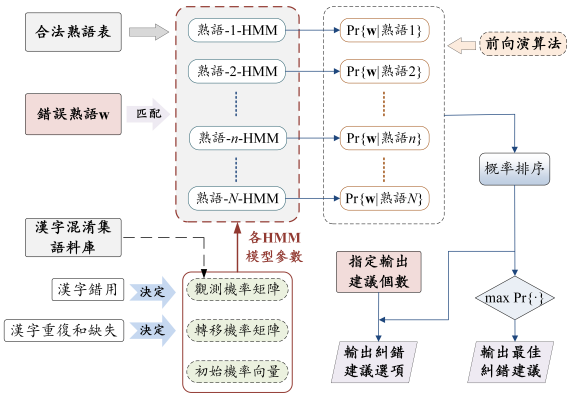

Figure 3: 基於 HMM 模型的中文熟語自動糾錯方法實施方案

餘下的問題是如何爲每個合法熟語確定 HMM 模型參數 $\{N_n, M_n, \mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n\,\forall n\}$。爲此，需要重新思考熟語使用錯誤的產生機理，進而確定建模方法。

首先，我們將隱藏狀態個數 N_n 設置爲合法熟語 $\mathbf{c}_n$ 的長度（即 $\mathbf{c}_n$ 包含的字元數），將觀測狀態個數 M_n 設置爲合法漢字的個數。例如，合法熟語「天南地北」的 HMM 模型含有 $N_n = 4$ 個隱藏狀態，分別爲「天」、「南」、「地」、「北」；考慮漢字 Unicode 編碼區間 0x4E00 至 0x9FA5，共有 20902 個合法漢字，對應 $M_n = 20902$ 個觀測狀態。

其次，我們考慮三種熟語錯誤來源，即漢字重複、缺失和錯字。例如，將「天南地北」誤用爲「天南地地北」，是漢字重複錯誤；將「天南地北」誤用爲「天南北」，是漢字缺失錯誤；將「天南地北」誤用爲「田南地北」，是漢字錯字錯誤。[2] 進一步地，重複錯誤和缺失錯誤使用轉移機率矩陣 $\mathbf{A}_n$ 表示，錯字錯誤則使用混淆機率矩陣 $\mathbf{B}_n$ 表示。

具體而言，$[\mathbf{A}_n]_{i,i}$ 表示合法熟語 $\mathbf{c}_n$ 的第 i 個字元發生重複錯誤的機率，$[\mathbf{A}_n]_{i,i+1}$ 表示合法熟語 $\mathbf{c}_n$ 的第 i 個字元正確地轉移到第 $i+1$ 個字元的機率，而 $[\mathbf{A}_n]_{i,i+k}$ 表示合法熟語 $\mathbf{c}_n$ 的第 $i+1$ 至 $i+k-1$ 個字元（$k>1$）發生缺失錯誤的機率；$[\mathbf{B}_n]_{i,j}$ 則表示合法熟語 $\mathbf{c}_n$ 的第 i 個字元被使用爲第 j 個合法漢字的機率。當然，假設合法熟語 $\mathbf{c}_n$ 的第 i 個字元是合法漢字集合中的第 i' 個漢字，則 $[\mathbf{B}_n]_{i,i'}$ 是該字

[2]值得說明的是，本文所提出的自動糾錯方法，對於重複、缺失和錯字等錯誤出現不止一次，甚至多種錯誤同時出現的複雜情況（如將「天南地北」誤用爲「天天南地地北」、「田楠帝北」、「天南帝北北北」等），都能夠予以糾正。詳見第4.1節「基本功能的檢驗和展示」。

元未出現錯誤的正確機率。矩陣 $\mathbf{A}_n$ 和 $\mathbf{B}_n$ 的行和（row-sum）都爲 1。

需要強調的是，爲了得到 $\mathbf{A}_n$、$\mathbf{B}_n$ 和 $\boldsymbol{\pi}_n$，理論上需要對熟語 $\mathbf{c}_n$ 的正確及錯誤使用情況進行大量統計。然而，大規模訓練資料的獲取難度極大。事實上，可以預期的是，即使僅對 $\mathbf{A}_n$、$\mathbf{B}_n$ 和 $\boldsymbol{\pi}_n$ 所有元素作合理的手工設置，本方法仍能表現出優秀的自動糾錯性能。我們將在第3.2節詳細說明各合法熟語 HMM 模型參數的賦值思路，並在第4章中結合實驗結果展開討論。本文所提出基於 HMM 模型的中文熟語自動糾錯實施方案如圖3所示。

3 實驗設置

3.1 資料集

本方法在構建自動糾錯模型時，需要爲每一個合法熟語建立各自的 HMM 模型，進而計算各 HMM 模型產生待糾錯熟語的後驗機率，最終根據機率值的排序給出糾正建議。因此，需要構建合法熟語表（作爲糾錯參考）、漢字混淆集（涵蓋近形字、同音字、近音字等，用於 HMM 模型參數設置）、熟語錯誤案例集（用於測試方法性能）等。具體如下：

1) 在構建合法熟語表時，我們使用 THUOCL 資料集 (Han et al., 2016) 中的成語詞庫作爲合法熟語表，如圖4所示。THUOCL 詞庫是由 Zhiyuan Liu 團隊整理推出的中文詞庫。該成語詞庫包含 8519 個成語，具有較高的代表性。

2) 在構建漢字混淆集時，已有 nlp-hanzi-similar 近形字[3]和 SimilarCharacter 同音字語料庫[4]等資料可供使用。此外，我們還自行構建了近音字語料庫。在構建近音字語料庫過程中，需用到「漢字轉拼音」和「拼音轉所有漢字」的基本操作，而該任務可以基於開放漢語字典-現代漢語字音資料庫[5]來實現。

3) 在測試本方法性能時，需要搜集具有代表性的熟語使用錯誤案例集。現有翰霖文教機構[6]、上海外國語大學[7]、揚州大學[8]、上海外國語大學附屬浦東外國語學校[9]等整理的資料，以及《多功能實用成語典》(Cai, 2016) 和《成語糾錯手冊》(Gao and Liu, 2011) 等著作可供使用。

[3] https://github.com/houbb/nlp-hanzi-similar
[4] https://github.com/contr4l/SimilarCharacter
[5] https://github.com/kfcd/hyzd
[6] https://www.han-lin.tw/chinese-form/
[7] http://www.newoaa.shisu.edu.cn/cc/cf/c6349a117967/page.htm
[8] https://jwc.yzu.edu.cn/info/1054/1902.htm
[9] https://www.msshw.pudong-edu.sh.cn/list/36/11580.html

Figure 4: THUOCL 中文詞庫，截圖自：http://thuocl.thunlp.org/

3.2 HMM 模型參數設置

如前文所述，理論上需要對各合法熟語 $\{\mathbf{c}_n\,\forall n\}$ 的使用情況進行大量統計，從而訓練出各合法熟語 HMM 模型參數 $\{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n\,\forall n\}$。不過，此操作需要消耗大量資源，現實中很難實現。可以預期的是，即使僅對 $\{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n\,\forall n\}$ 進行簡單合理的手工設置，本方法仍能表現出優秀性能。本節將介紹具體的設置思路和方法。

3.2.1 轉移機率矩陣 $\{\mathbf{A}_n\,\forall n\}$

如無其他先驗知識，在本文中，長度相等的合法熟語使用相同的轉移機率矩陣。例如，對於長度爲 $N = 6$ 的合法熟語「有志者事竟成」，其轉移機率矩陣 $\mathbf{A}$ 可簡單設置爲

$$\mathbf{A} = \begin{bmatrix} 0.10 & 0.80 & 0.07 & 0.03 & 0 & 0 \\ 0 & 0.10 & 0.80 & 0.07 & 0.03 & 0 \\ 0 & 0 & 0.10 & 0.80 & 0.07 & 0.03 \\ 0 & 0 & 0 & 0.10 & 0.80 & 0.10 \\ 0 & 0 & 0 & 0 & 0.10 & 0.90 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

在式 (3) 中，矩陣 $\mathbf{A}$ 由一個基本機率質量函數（probability mass function，PMF）進行移位（shifted）而產生。我們將該基本 PMF 記爲 $\mathbf{a}_0$。因此，式 (3) 中的 $\mathbf{a}_0$ 是一個有效長度爲 $L_{\mathbf{A}} = 4$ 的行向量（row vector），其取值爲

$$\mathbf{a}_0 = [0.10, 0.80, 0.07, 0.03]. \tag{4}$$

式 (4) 表示該合法熟語每個漢字的「正確轉移機率」都爲 0.80，漢字重複機率都爲 0.10，缺失 1 個漢字的機率爲 0.07，缺失 2 個漢字的機率爲 0.03，而缺失更多漢字的機率均爲

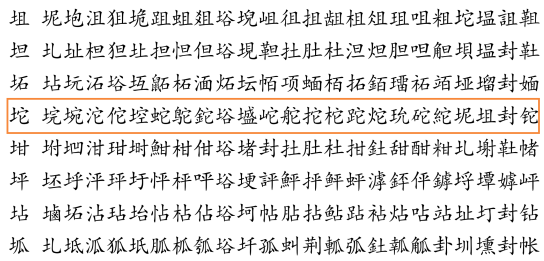

Figure 5: nlp-hanzi-similar 語料庫 (Hou, 2025) 的部分資料（該語料庫用於對近形字錯誤進行建模），整理自：https://github.com/houbb/nlp-hanzi-similar

Figure 6: SimilarCharacter 語料庫 (XiaoFang et al., 2025) 的部分資料（該語料庫用於對同音字錯誤進行建模），截圖自：https://github.com/contr4l/SimilarCharacter

0。[10] 同時，矩陣 $\mathbf{A}$ 的下對角線元素（entries strictly below the main diagonal）均設爲 0，這是由於假設使用該合法熟語時發生漢字位置調換的機率小到可以忽略。當然，使用者可以根據實際情況，對矩陣 $\mathbf{A}$ 的基本 PMF 進行修改或學習。

3.2.2 觀測機率矩陣 $\{\mathbf{B}_n \forall n\}$

觀測機率矩陣 $\mathbf{B}_n$ 各元素的值表示各合法漢字錯用爲其它合法漢字的機率。同樣地，理論上需要利用漢字錯用資料進行統計學習，但這需要大量的計算資源，且相關資料並不充分，關於漢字錯用類型的分類也尚無統一的規範 (Liu et al., 2008)，現有資料難以直接利用。

鑒於此，本文提出一種簡單有效的觀測機率矩陣設置方法。首先，利用現有語料庫，爲每個合法漢字找到其對應的所有近形字、同音字和近音字。需要說明的是，「音」是指拼音，而在演算法實現時，可根據實際需求使用相應的拼音系統，例如漢語拼音、注音符號第二式、威妥瑪拼音、通用拼音等。本文以漢語拼音爲例進行闡述和實現。進而，將錯用機率在其近形字、同音字、近音字和所有其它合法漢字中進行分配。具體操作如下：

1) 由近形字導致的錯用情況，使用 nlp-hanzi-similar 語料庫 (Hou, 2025) 進行設置。例如，在圖5中，對於漢字「坨」，考慮「垸」、「堍」、「沱」、「佗」等近形字。

2) 由同音字導致的錯用情況，使用 SimilarCharacter 語料庫 (XiaoFang et al., 2025) 進行設置。例如，在圖6第 1599 行中，對於漢字「乖」，考慮「拐」、「怪」、「捆」共 3 個同音字[11]。

3) 對於由近音字導致的錯用情況，我們構建了近音字語料庫。在本文已完成的簡單原型系統中，暫時只考慮由前後鼻音以及平翹舌音導致的近音字錯誤。例如，拼音爲「can」的漢字，其近音字包括拼音爲「cang」和「chan」的所有漢字；拼音爲「an」的漢字，其近音字僅包括拼音爲「ang」的所有漢字。[12] 對此，我們需要獲取每個漢字的拼音以及獲取各相近拼音對應的所有漢字，即實現「漢字轉拼音」和「拼音轉所有漢字」。

最後，結合上述近形字、同音字和近音字錯誤資料，可對各合法漢字錯用爲其它合法漢字的觀測機率值進行設置。令漢字使用的正確率爲 $p_{\mathbf{B}}$，則需要將錯用機率 $1-p_{\mathbf{B}}$ 在除該漢字之外的所有合法漢字中進行分配。假設近形字錯誤、同音字錯誤、近音字錯誤和其它錯誤等共計 4 種錯誤所佔比例分別爲 x_1、x_2、x_3 和 x_4，滿足 $x_1+x_2+x_3+x_4=100\%$。若無其它先驗知識，一種簡單有效的做法是：將近形字錯誤機率 $(1-p_{\mathbf{B}})x_1$、同音字錯誤機率 $(1-p_{\mathbf{B}})x_2$、近音字錯誤機率 $(1-p_{\mathbf{B}})x_3$ 和其它錯誤機率 $(1-p_{\mathbf{B}})x_4$ 分別在所對應的所有合法漢字中平均分配，最終得到觀測機率矩陣各元素的值。例如，當 $p_{\mathbf{B}}=0.90$，且 $x_1=20\%$、$x_2=40\%$、$x_3=30\%$ 和 $x_4=10\%$ 時，將近形字錯誤機率 $0.10x_1=0.02$、同音字錯誤機率 $0.10x_2=0.04$、近音字錯誤機率 $0.10x_3=0.03$ 和其它錯誤機率 $0.10x_4=0.01$ 分別在對應漢字中平均分配。

[10] 矩陣 $\mathbf{A}$ 的第 $N-L_{\mathbf{A}}+1$ 至 N 行的上對角線元素（entries strictly above the main diagonal）需要進行修正，以保證行和（row-sum）均爲 1，如式 (3) 中的第 3 至 6 行所示。

[11] 在本文中，爲方便起見，「同音字錯誤」是指拼音相同但音調可能不同的情形（如「廣」和「逛」也視爲同音字），並不局限於拼音和音調都相同。

[12] 後續工作中，使用者可以根據實際情況，對近音字錯誤類型進行修改和擴充，例如定義「nan」和「lan」爲相近拼音。

錯誤熟語	錯誤類型及個數	**Word** 檢查器	**GPT-5**	本方法
風中之濁	近形 1	風中之燭	風中之燭	風中之燭
逢場做戲	同音 1	逢場作戲	逢場作戲	逢場作戲
發號司令	近音 1	無糾錯建議	發號施令	發號施令
爭先後	缺字 1	無糾錯建議	爭先恐後	爭先恐後
更新換換代代	多餘 2	無糾錯建議	更新換代	更新換代
屋裡去鬧	同音 3	無糾錯建議	無理取鬧	無理取鬧
瘋馬牛兒不想記	同音 3、多餘 1	無糾錯建議	風馬牛不相及	風馬牛不相及
掛狗頭買羊肉	錯字 2、同音 1	無糾錯建議	掛羊頭賣狗肉	掛羊頭賣狗肉
畢奇工夫於一夜	同音 2、多餘 1、錯字 1	無糾錯建議	畢其功於一役	畢其功於一役
意點異低	同音 3	無糾錯建議	疑點重重	一點一滴
天婚棣岸	同音 3	無糾錯建議	天崩地裂	天昏地暗
有四五孔	同音 2、近音 1	無糾錯建議	有恃無恐	有恃無恐
溜繩吳組	同音 2、近音 2	無糾錯建議	無功受祿	六神無主
億民金仍	同音 2、近音 2	無糾錯建議	一鳴驚人	一鳴驚人
拔山涉稅	同音 2	無糾錯建議	拔山涉水	跋山涉水
图挠蜈蚣	同音 3、錯字 1	無糾錯建議	獨木難支	徒勞無功
一洋囊進	同音 1、近音 2	無糾錯建議	一網打盡	一言難盡
澀澀森威	同音 2、近音 1	無糾錯建議	虎虎生威	虎虎生威
喜夾良李	錯字 1、近音 1	無糾錯建議	喜出望外	喜結連理
恭喜花柴	錯字 1、近音 1	無糾錯建議	恭喜發財	恭喜發財

Table 1: 不同出錯嚴重程度下的中文熟語糾錯結果對比示例

測試集	樣本數	準確 糾正數	糾正 準確率
G&L	1285	1241	96.6%
YZU	104	101	97.1%
SISU	47	46	97.9%
SISU-PD	183	175	95.6%
總計	1619	1563	96.5%

Table 2: 使用測試集對本方法糾錯能力進行評估

3.2.3 初始機率向量 $\{\boldsymbol{\pi}_n \forall n\}$

在本文的建模思路中，$\boldsymbol{\pi}_n$ 的第 i 個元素表示人們在使用合法熟語 $\mathbf{c}_n$ 時，首個漢字是 $\mathbf{c}_n$ 的第 i 個漢字的機率。通常情況下，$\boldsymbol{\pi}_n$ 的第 1 個元素應相對較大，即正確機率相對較大。在後續實驗中，如無其他先驗知識，我們將 $\boldsymbol{\pi}_n$ 的第 1 個元素都設置爲 $[\boldsymbol{\pi}_n]_{1,1} = \pi_1$，其餘第 $j\,(j>1)$ 個元素設置爲 $[\boldsymbol{\pi}_n]_{1,j} = \frac{1-\pi_1}{N_n-1}$。

3.3 評估準則

我們首先在第4.1節完成基本功能測試實驗，然後在第4.2和4.3節使用測試集進行全面評估。測試集包含多個測試樣本，每個測試樣本包含 1 個「錯誤寫法」和 1 個「正確寫法」。在測試實驗中，以每個測試樣本中的「錯誤寫法」作爲輸入，取本方法給出的機率最大的 N_{cand} 個糾錯建議作爲輸出，並與測試樣本中的「正確寫法」作比較。若至少 1 個糾錯建議與「正確寫法」相同，則視爲「準確糾正」，否則視爲「錯誤糾正」。統計各測試集的準確糾正數和錯誤糾正數，以準確糾正數與測試樣本數之比爲性能指標，稱糾正準確率（correction rate，CR）。

4 實驗結果及討論

4.1 基本功能的檢驗和展示

在本節中，我們以出錯程度不同的熟語作爲測試輸入，取本方法所給出的最佳糾錯建議作爲測試輸出（$N_{\text{cand}} = 1$），與微軟公司 Word 辦公軟體拼寫檢查器以及 GPT-5 的智能糾正結果進行比較，以檢驗本方法的基本功能。我們在不同測試案例中加入不同數量的近形字、同音字、近音字、缺失、多餘等錯誤類型，甚至有多種錯誤同時出現的情況。

在實驗中，所有轉移機率矩陣的基本 PMF 都設置爲 $\mathbf{a}_0 = [0.10, 0.80, 0.07, 0.03]$，漢字正確使用機率 $p_{\mathbf{B}} = 0.90$，$x_1 = 20\%$，$x_2 = 40\%$，$x_3 = 30\%$，$x_4 = 10\%$，初始機率向量 $\pi_1 = 0.9$，$\boldsymbol{\pi}_n = [\pi_1, \frac{1-\pi_1}{N_n-1}, \cdots, \frac{1-\pi_1}{N_n-1}]$。實驗結果如表1所示。[13]

[13]對於 GPT-5，所使用的提示詞爲：「風中之濁、逢場做戲、發號司令、爭先後、更新換換代代、屋裡去鬧、瘋馬牛兒不想記、掛狗頭買羊肉、畢奇工夫於一夜、意點異低、天婚棣岸、有四五孔、溜繩吳組、億民

序號	測試樣本中的錯誤寫法	本方法糾正建議	測試樣本中的正確寫法	說明或啓示
1	固步自封	固步自封	故步自封	「固步自封」並非公認的正確用法。合法熟語表的專業性、嚴謹性和無歧義性十分重要。
2	流言非語	流言飛語	流言蜚語	「流言飛語」和「流言蜚語」是異形詞。合法熟語表的專業性、嚴謹性和無歧義性十分重要。
3	辛辛學子	過河卒子	莘莘學子	後續在完善漢字混淆集時，需對易錯字情況進行更全面的考慮。
4	發人深醒	發人深思	發人深省	後續在完善漢字混淆集時，需對一字多音的情況加以考慮。

Table 3: 表2中 SISU 和 YZU 兩個測試集的 4 個錯誤糾正測試樣本的細節及討論

由表1可見，Word 拼寫檢查器的糾錯能力最差。如前文所述，這是因爲該檢查器使用了自定義詞典，只能糾正已收錄於詞典中的特定錯誤（例如「風中之燭」和「逢場做戲」這兩個錯誤詞），而對於其它錯誤詞則無法處理。與之不同的是，本方法發揮了 HMM 模型在序列分析和機率建模方面的優勢，能夠靈活地糾正多樣化類型的熟語使用錯誤。

有趣的是，GPT-5 無法對一些錯誤熟語予以正確糾正，尤其是同音錯誤和近音錯誤比較嚴重時。例如，在表1中，GPT-5 建議將「意點異低」糾正爲「疑點重重」而非「一點一滴」，將「天婚棣岸」糾正爲「天崩地裂」而非「天昏地暗」，將「一洋囊進」糾正爲「一網打盡」而非「一言難盡」。同時，GPT-5 有時會給出錯誤的糾正建議，即糾正結果並不是公認的合法熟語，例如表1中的「拔山涉水」。相反，本方法模型簡單、參數少，無需複雜的訓練過程，仍能表現出與大語言模型相當（甚至更優）的熟語糾錯性能。當然，我們的目的不在於貶低 GPT-5 的熟語糾錯能力（畢竟熟語糾錯並非其唯一關注點），而是將其作爲對比基準之一，驗證本方法的有效性和靈活性。

4.2 使用測試集對系統性能進行評估

本節評估本方法的糾正准確率。採用鳳凰出版社《成語糾錯手冊》、揚州大學、上海外國語大學和上海外國語大學附屬浦東外國語學校等整理發佈的共計四個常見成語使用錯誤案例集作爲測試集，下文分別稱爲 G&L、YZU、SISU 和 SISU-PD 測試集。這四個測試集分別有 1285、104、47 和 183 個測試樣本，共計 1619 個。每個測試樣本包含 1 個「錯誤寫法」和 1 個「正確寫法」，例如錯誤的「直接了當」和正確的「直截了當」。同時，在測試實驗中，若測試樣本的「正確寫法」事先未收錄於 THUOCL 成語庫中，則將其加入並更新合法成語表，以完善合法成語表的收錄質量。

在實驗中，轉移機率矩陣基本 PMF（即 $\mathbf{a}_0$）、漢字正確使用機率 $p_{\mathbf{B}}$、漢字錯誤使用機率分配比例 (x_1, x_2, x_3, x_4)、初始機率向量 $\boldsymbol{\pi}_n$、糾正建議個數 N_{cand} 的設置都與第4.1節相同。實驗結果如表2所示。由表2可見，即使僅對 $\{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n \forall n\}$ 進行簡單合理的手工設置，本方法對四個測試集的糾正準確率分別高達 96.6%、97.1%、97.9% 和 95.6%，總計 96.5%，因而在統計意義上具有優秀的糾錯能力。這說明本方法對熟語出錯過程的建模，以及對合法熟語 HMM 模型的轉移機率和觀測機率等參數的設置及相應物理意義的理解，都是合理且有效的。

同時，爲了更好地理解本方法的特點，我們將表2中 SISU 和 YZU 兩個測試集的 4 個錯誤糾正測試樣本記錄下來，如表3所示，並作討論如下：

1) 第 1 個樣本中，本方法認爲「固步自封」一詞無錯誤。該測試樣本之所以未準確糾正，是由於實驗所使用的 THUOCL 成語表收錄了「固步自封」一詞。因此，該錯誤糾正與本方法的核心技術無關。當然，這也給本文後續工作帶來啓示：合法熟語表的專業性、嚴謹性和無歧義性十分重要。

2) 第 2 個樣本中，本方法將「流言非語」糾正爲「流言飛語」，而測試樣本的正確寫法是「流言蜚語」。經查閱資料，「流言飛語」和「流言蜚語」是一組異形詞 (Wang et al., 2001)。然而，THUOCL 成語表收錄了「流言飛語」一詞。因此，與上一個樣本相同，該錯誤糾正與本方法的核心技術無關。

3) 第 3 個樣本中，本方法未將「辛辛學子」糾正爲「莘莘學子」。這是因爲在構建漢字混淆集時，未充分考慮「莘」字容易被錯用爲「辛」字這一情況，即未將「莘」和「辛」視爲

金仍、拔山涉稅、图挠蜈蚣、一洋囊進、滸滸森威、一萬无寄、恭喜花柴，以上這些詞，分別最可能是哪個熟語的錯誤使用？」網址爲：https://chatgpt.com/。

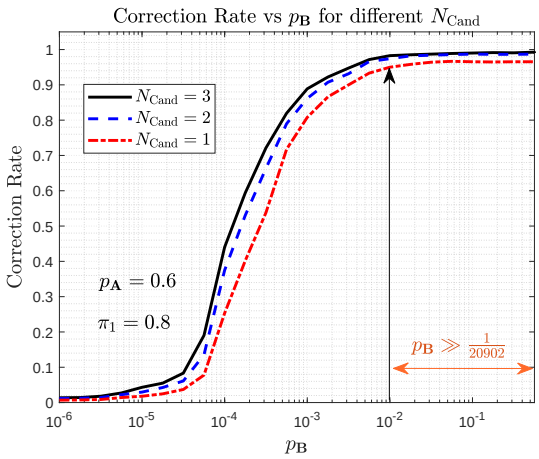

Figure 7: 正確糾正率受觀測機率矩陣取值影響的實驗結果圖，其中 N_{cand} 表示本方法輸出的糾正建議個數

近形字或近音字。儘管這個錯誤糾正與本方法的核心建模思路無關，但也啓示我們：後續需對易錯字情況進行更全面的考慮，進一步完善漢字混淆集的構建。

4) 第 4 個樣本中，本方法未將「發人深醒」糾正爲「發人深省」。這是因爲在構建同音字混淆集時，只考慮「省」字的讀音「sheng」，未考慮「省」字的另一個讀音「xing」，即未將「省」和「醒」視爲同音字。儘管這個錯誤糾正與本方法的核心建模思路無關，但也啓示我們：後續需對一字多音的情況加以考慮，進一步完善漢字混淆集的構建。

4.3 健壯性測試

在本文中，我們並未嘗試通過收集每一個合法熟語 $\{\mathbf{c}_n \forall n\}$ 的正確及錯誤使用情況（因其代價巨大）來訓練每一個 HMM 模型的參數 $\{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n \forall n\}$。有趣的是，由第4.1和4.2節可見，本方法只需對 $\{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n \forall n\}$ 進行簡單合理的手工設置，就能得到良好的糾正準確率。事實上，由於使用了機率模型而非編輯距離模型，並且各合法熟語之間天然具有較明顯的差別，本方法對於 $\{\mathbf{A}_n, \mathbf{B}_n, \boldsymbol{\pi}_n \forall n\}$ 的不同取值具有很強的健壯性（robustness）。對此，我們完成以下實驗進行驗證。

實驗採用第4.2節中的四個測試集共計 1619 個測試樣本進行。同時，對於長度爲 N_n 的合法熟語 $\mathbf{c}_n$，其對應的 HMM 模型參數設置如下：狀態轉移矩陣 $\mathbf{A}_n$ 的基本 PMF 爲 $\mathbf{a}_0 = [\frac{1-p_{\mathrm{A}}}{3}, p_{\mathrm{A}}, \frac{1-p_{\mathrm{A}}}{3}, \frac{1-p_{\mathrm{A}}}{3}]$；混淆機率矩陣 $\mathbf{B}_n$ 中，$x_1 = x_2 = x_3 = 30\%$，$x_4 = 10\%$；初始機率向量設置爲 $\boldsymbol{\pi}_n = [\pi_1, \frac{1-\pi_1}{N_n-1}, \cdots, \frac{1-\pi_1}{N_n-1}]$。

首先，我們考察觀測機率矩陣 $\{\mathbf{B}_n \forall n\}$ 中的漢字正確使用機率 p_{B} 對糾正準確率的影響，實驗結果如圖7所示。由圖可見，只要 p_{B} 取值明顯大於 $\frac{1}{20902}$，本方法都能給出高於 95% 的糾正准確率，呈現出關於觀測機率矩陣的健壯性。[14] 另外，糾正准確率隨著輸出建議個數 N_{cand} 的增加而提升。有趣的是，即便只取機率最大的唯一一個糾錯建議作爲輸出（$N_{\mathrm{cand}} = 1$），在 $p_{\mathrm{B}} > 0.01$ 時也能獲得高於 95% 的糾正准確率。最後，我們考察轉移機率矩陣基本 PMF 中的 p_{A} 和初始機率向量的 π_1 值對糾正准確率的影響，實驗結果如表4和5所示。可以看出，本方法對於 $\{\mathbf{A}_n \forall n\}$ 和 $\{\boldsymbol{\pi}_n \forall n\}$ 的不同取值同樣具有很強的健壯性。以上特點有助於降低收集大量熟語使用情況進行模型參數學習的必要性。

p_{A}	0.4	0.6	0.8	1
$\mathrm{CR}_{0.01}$	94.6%	94.9%	95.3%	94.3%
$\mathrm{CR}_{0.1}$	96.7%	96.7%	96.9%	95.4%

Table 4: 糾正準確率受轉移機率矩陣取值影響的實驗結果，其中 CR_x 表示 $p_{\mathrm{B}} = x$ 時的結果

π_1	0.4	0.6	0.8	1
$\mathrm{CR}_{0.01}$	95.1%	95.3%	95.2%	95.1%
$\mathrm{CR}_{0.1}$	96.9%	96.9%	96.8%	96.6%

Table 5: 糾正準確率受初始機率向量取值影響的實驗結果，其中 CR_x 表示 $p_{\mathrm{B}} = x$ 時的結果

5 結論

本文提出了一種輕量型的中文熟語自動糾錯新方法。通過構建包含近形字、同音字和近音字的漢字混淆集，以及利用現有的合法術語庫，將中文熟語糾錯問題建模爲隱藏式馬可夫模型的基本問題並求解。相比於傳統的編輯距離和自定義詞典方法，本方法模型參數少、計算簡單，即使僅對模型參數進行簡單合理的人工設置，也能獲得出色的糾錯性能，對出錯程度嚴重的熟語予以準確糾正。同時，本方法對模型參數具有很強的健壯性，因而無需耗費大量資源用於訓練資料獲取和模型參數學習。在後續的工作中，需要完善合法熟語表和漢字混淆集的構建，以提升模型的糾正準確性。此外，可進一步研究如何推廣到考慮上下文資訊的場景，以增大本方法的適用范圍。

[14] $p_{\mathrm{B}} > \frac{1}{20902}$ 在實際使用中很容易滿足。此處的 20902 是觀測狀態個數，即合法漢字個數。此外，p_{B} 不能嚴格等於 1，因爲 $p_{\mathrm{B}} = 1$ 會導致 $\Pr\{\mathbf{w}|\lambda_n\} = 0 \forall n$，從而導致異常結果。

References

Zongyang Cai. 2016. 多功能實用成語典（第二版）*(Multi-functional Practical Idioms Dictionary (Second Edition)) [in Chinese]*. Taipei: Wu-Nan Book Inc.

Yulin Gao and Peishu Liu. 2011. 成語糾錯手冊 *(Idiom Correction Manual) [in Chinese]*. Nanjing: Phoenix Publishing and Media Inc.

Shiyi Han, Yuhui Zhang, Yunshan Ma, Cunchao Tu, Zhipeng Guo, Zhiyuan Liu, and Maosun Sun. 2016. THUOCL：Tsinghua Open Chinese Lexicon. Website. http://thuocl.thunlp.org/.

Victoria J. Hodge and Jim Austin. 2003. A comparison of standard spell checking algorithms and a novel binary neural approach. *IEEE Transactions on Knowledge and Data Engineering*, 15(5):1073–1081.

Binbin Hou. 2025. 漢字相似度計算工具及中文形近字算法 (Chinese character similarity calculation tools and algorithms for Chinese similar-looking characters) [in Chinese]. Website. https://github.com/houbb/nlp-hanzi-similar.

Amita Jain and Minni Jain. 2014. Detection and correction of non word spelling errors in Hindi language. In *2014 International Conference on Data Mining and Intelligent Computing (ICDMIC)*, pages 1–5.

Dan Jurafsky and James H. Martin. 2024. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (Third edition draft). Website. https://web.stanford.edu/~jurafsky/slp3/.

Muhammad Ifte Khairul Islam, Rahnuma Islam Meem, Faisal Bin Abul Kasem, Aniruddha Rakshit, and Md. Tarek Habib. 2019. Bangla spell checking and correction using edit distance. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–4.

Weihui Lin, Weijun Tang, Xiaofan Lin, Rongbin Zhang, Shaojie Xu, Xiaojuan Ning, and Wenting Zhang. 2012. *Hidden Markov model and its application in typewriting correction.* Project report in the course Machine Learning and Its Applications, South China University of Technology, Guangzhou.

Chao-Lin Liu, Chih-Bin Huang, Juei-Yu Weng, and Yi-Hsuan Chuang. 2008. 形音相近的易混淆漢字的搜尋與應用 (Identification and applications of visually confusing Chinese characters) [in Chinese]. In *Proceedings of the 20th Conference on Computational Linguistics and Speech Processing (ROCLING)*, pages 108–122.

Microsoft-Corporation. 2025. Add or edit words in a spell check dictionary. Website. https://support.microsoft.com/.

Ei Phyu Phyu Mon, Ye Kyaw Thu, Than Than Yu, and Aye Wai Oo. 2021. SymSpell4Burmese: Symmetric delete spelling correction algorithm (SymSpell) for Burmese spelling checking. In *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6.

Peter Norvig. 2016. How to write a spelling corrector. Website. http://norvig.com/spell-correct.html.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

A. Revathi, M. Vimaladevi, and N. Arivazhagan. 2023. Spelling correction using encoder-decoder and Damerau-Levenshtein distance. In *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, pages 469–472.

Moch Yusup Soleh and Ayu Purwarianti. 2011. A non word error spell checker for Indonesian using morphologically analyzer and HMM. In *Proceedings of the International Conference on Electrical Engineering and Informatics*, pages 1–6.

Mark Stamp. 2021. A revealing introduction to hidden Markov models. Website. https://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf.

Kseniia Varlamova, Ildar Khabutdinov, and Andrey Grabovoy. 2023. Automatic spelling correction for Russian: Multiple error approach. In *2023 Ivannikov Ispras Open Conference (ISPRAS)*, pages 169–175.

Jihong Wang, Ming Chen, and Liqing Ren. 2001. 現代實用漢語字典 *(Modern Practical Chinese Dictionary) [in Chinese]*. Shanghai: Shanghai Far East Publishers.

Ziqi Wang, Gu Xu, Hang Li, and Ming Zhang. 2014. A probabilistic approach to string transformation. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1063–1075.

XiaoFang, mystical001, and demo. 2025. 對常用的 6700 個漢字進行音、形比較，輸出音近字、形近字的列表 (Compare the sounds and shapes of 6,700 commonly used Chinese characters and output a list of characters with similar sounds and shapes) [in Chinese]. Website. https://github.com/contr4l/SimilarCharacter.

Chengqing Zong. 2024. Lecture Notes on Natural Language Understanding, Chinese Academy of Sciences, Beijing. Website. https://nlpr.ia.ac.cn/cip/ZongReportandLecture/ReportandLectureIndex.htm.