漢語間統計式機器翻譯語料處理 - 臺灣閩南語示範

Corpus Preprocessing for Statistical Machine Translation between the Chinese Languages - Using Taiwan Southern Min as an Example

103/11/06
國立交通大學 資訊工程與科學研究所
0156016 薛丞宏
指導教授：張智星教授
易志偉教授

目錄

第一節：研究背景

- 臺灣是多元語言的國家
  - 南島語
    - 阿美、泰雅、噶哈巫、西拉雅、...
  - 漢語
    - 閩南語、客話、華語（官話）、二戰後移民...
  - 其他
    - 越南（新住民）、...
- 講母語是人上基本的權利
  - 毋過臺灣母語消失誠緊
  - 學習資源無夠

研究方向

- 共華語翻譯做母語
  - 華語資料誠濟
  - 增加母語資源
    - 予欲學的人參考
  - 配合語音合成

- 華語閩南語的翻譯
  - 主要處理翻譯語料
    - 無修改翻譯演算法
  - 結果嘛會使用佇客話

第二節：相關文獻佮背景智識

- **語料**
  - 語料狀況
  - 語料庫介紹
  - 語料樣式
- 翻譯相關
  - 斷詞方法
  - 翻譯模型
  - 語料整理

閩南語語料種類

種類	範例	語料	備註
全漢	我欲食飯		全部漢字
全羅	gua2 beh4 tsiah8-png7		全部羅馬拼音 有斷詞資訊
漢羅	我beh4食飯	TGB通訊	漢字拼音濫咧用
全漢全羅對應	我欲食飯 gua2 beh4 tsiah8-png7	新聞語料庫 教育部辭典	漢字佮拼音攏有
漢羅全羅對應	我beh4食飯 gua2 beh4 tsiah8-png7	臺語文數位典藏	部份字有漢字

- 逐个語料庫攏無仝
  - 後壁愛處理的問題

語料庫－教育部辭典

- 全名「臺灣閩南語常用詞辭典」
  - 較濟生活用語
- 有全漢、全羅佮斷詞
- 詞條
  - 116724詞
- 例句有華語翻譯
  - 8027句
- 附錄句無華語翻譯
  - 388句

全漢	彼个查某囡仔真媠。
全羅	hit4 e5 tsa-boo2 gin2-a2 tsin1 sui2.
華語	那個女孩子很漂亮。

語料庫－新聞語料庫

- 全名「臺華平行新聞語料庫」
  - 中研院資訊所陳孟彰老師主持，何澤政翻譯
  - 翻譯時，罕得調整語詞先後
  - 現代用語、古早用語攏有
- 有全漢、全羅，斷詞無規範
- 97/11/06到103/3/14的文章
  - 2567篇文章、64121句
  - 359554華語詞組、366190閩南語詞組

全漢	這幾工 寒流 閣再 展威
全羅	tsit4-kui2-kang1 han5-liu5 koh4-tsai3 tian2-ui1
華語	這幾天 寒流 再度 發威

語料庫－臺文典藏

- 全名「台語文數位典藏」
  - 國家臺灣文學館收集1885～2006年的語料
- 有漢羅、全羅
  - 臺文館後來倩人拍字，補資料
  - 有的劇本全羅內底有漢字
- 攏總2167篇
  - 詩387條
  - 散文1127篇
  - 小說387篇
  - 劇本49篇

漢羅	Koh m7知u7危險
全羅	Koh m7-tsai u7 gui5-hiam2

語料庫－TGB通訊

- 「學生台灣語文促進會」主持
  - 一個月一期的部落格
- 有漢羅佮華語
  - 無一定是平行語料
- 1999年到這馬
  - 實驗的資料到2014年6月12日，1179篇文章

平行語料	範例
閩南語漢羅	Gín-á beh講siáⁿ-mih款語言kám是gín-á ka-tī決定--ê? Lah是大人提供ê選擇?
華語漢字	孩子要講什麼語言是孩子自己決定的嗎？還是大人提供的選擇？

腔口無仝

- 閩南語有許多腔調
  - 偏漳腔、混合腔、偏泉腔
  - 混和腔有較濟偏漳腔的特色
- 語料狀況
  - 教育部資料
    - 有地方腔，鹿港「火her2」
    - 主要資料有記錄腔口，附錄句無
  - 新聞語料庫
    - 澤政是臺中烏日人，60年代出身
    - 偏漳腔，有時陣會濫著泉腔
  - 臺文典藏
    - 四界收集來的，無記錄腔
  - TGB通訊
    - 無註明腔，干焦漢羅，無法度算
- 全部資料濫做伙訓練
  - 資料無逐个註明

教育部	偏漳	混合	偏泉	外來語
字/詞	36142	33756	46654	172
例句	10637	9829	14227	0

漳/泉	雞 ke1/kue1	近 kin7/kun7	火 hue2/he2
附錄句	16/0	5/0	2/0
新聞語料庫	284/13	710/45	1037/25
臺文典藏	229/135	458/365	703/390

語料樣式

- 斷字
  - 無斷詞
- 斷詞

語料樣式	範例
斷字	蔡崇名細漢時生活困苦
斷詞	蔡崇名 細漢 時 生活 困苦

- 後壁會比較這兩个對翻譯的影響

第二節目錄

- 語料
  - 語料狀況
  - 語料庫介紹
  - 語料樣式
- **翻譯相關**
  - 斷詞方法
  - 翻譯模型
  - 語料整理

斷詞方式

- 華語斷詞工具
  - 中研院中文斷詞系統（CKIP）
  - [1] Ma, Wei-Yun, 2003

- 閩南語斷詞工具
  - 長詞優先斷詞

長詞優先斷詞方法

- 長詞優先
  - 上定看著的斷詞方法
  - 對後壁開始看，希望詞愈長愈好
    - 華語實驗的結果，效果比對頭前閣較好
- 做法
  1. 對上後壁的字開始，看k个字
  2. 揣一个佇辭典的上長詞
  3. 揣著詞了後，繼續對第1步做到結束
- 方法
  - 若一句話攏總m字，$j_1, j_2, \dots, j_m$
  - 長詞優先斷詞(m)
    - 斷詞位置 $= \underset{i}{\text{argmin}}\{[j_{i+1}, j_{i+2}, \dots, j_m]$是一个詞$\}$
      $+$長詞優先斷詞$(i), m-k \le i \le m-1$

長詞優先斷詞範例

- 範例 - 蔡崇名細漢時生活困苦，k=5
  1. 蔡崇名細漢時生活困苦
  2. 蔡崇名細漢時生活困苦
  3. 蔡崇名細漢時生活困苦
  4. 蔡崇名細漢時生活困苦
  5. 蔡崇名細漢時生活困苦
  6. 蔡崇名細漢時生活 困苦
  7. 蔡崇名細漢時生活 困苦
  8. 蔡崇名細漢時生活 困苦
  9. 蔡崇名細漢時生活 困苦
  10. 蔡崇名細漢時 生活 困苦
  11. ……

斷詞評分方式

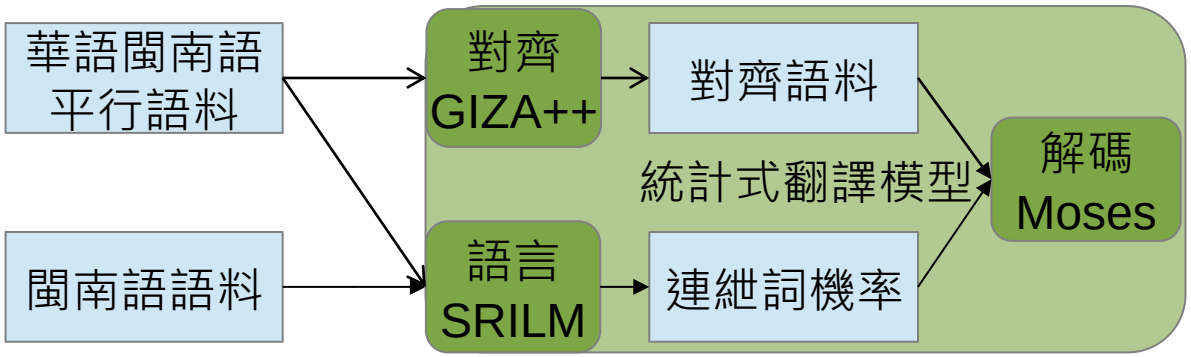
- 召回率 $=\frac{\text{斷著的斷詞數量}}{\text{答案的斷詞數量}}$
- 精確率 $=\frac{\text{斷著的斷詞數量}}{\text{結果的斷詞數量}}$
- $F-$測量 $=\frac{2\times\text{召回率}\times\text{精確率}}{\text{召回率}+\text{精確率}}$

翻譯模型

- 統計式翻譯
  - Brown et al., 1993
  - 對齊模型alignment model
    - 華語佮閩南語的詞對應機率
    - Och and Ney, 2003
  - 語言模型language model
    - 閩南語語句合理性
    - Stolcke, 2002
  - 解碼器decoder
    - 用對齊模型佮語言模型翻譯
    - Philipp Koehn et al. 2007.

華語	閩南語
他 早上 有來 **幫忙**	伊 早時 有來 **鬥相共**
幫忙 解決 問題	**鬥相共** 解決 問題

閩南語語句	合理程度
我 早時 **食 菜頭粿**	語料較會出現，機率懸
我 早時 **食 電話**	語料較罕得出現，機率低



翻譯評分方式 - BLEU

- Bilingual Evaluation Understudy
- 以連緌詞n-grams為單位
- $BLEU$

$$= 100 \times e^{\max(0,\frac{\text{結果}-\text{答案長度}}{\text{結果長度}})} (\prod_{n=1}^{4} \text{連緌詞}_n)^{\frac{1}{4}}$$

答案	這 幾 工 寒流 閣再 展威
結果一	這 幾 工 寒流 有 展威
結果二	寒流 這 幾 工 閣再 展威

連緌詞$_n$	n=1	n=2	n=3	n=4	BLEU
結果一	5/6	3/5	2/4	1/3	53.73
結果二	6/6	3/5	1/4	0/4	0.00

其他語料處理

- 語言分類language identification
  - 判斷一句話，是佗一个語言
  - Cavnar and Trenkle, 1994
    - 用語言模型算分數
    - 以字元為單位
- 平行語料語句對齊sentences alignment
  - 兩種語言的翻譯文章，對齊意思相倚的句
  - Sennrich and Volk, 2010
    - 用翻譯模型鬥對齊語料

貢獻

- 比較漢語語料樣式對翻譯的影響
- 分類兩種漢語的方法
- 提出一个整理漢語語料的方法
- 證明語料的數量對目前的成果影響誠濟

第三節：研究介紹

- 目標
  - 語料預處理，予華語閩南語翻譯，效果較好
    - 用BLEU做評分標準
- 語料形式愈統一翻譯愈好
  - 第一个問題，按怎斷詞（閩南語斷詞）
  - 第二个問題，有的詞翻袂出來（未知詞問題）
- 語料愈濟愈好，所以加語料
  - 第三个問題，按怎加入資料無完整的語料庫（語料整理）
  - 第四个問題，網路語料需要分華語佮閩南語（語言分類）

第一个問題 - 閩南語斷詞

- 輸入
  - 全漢佮全羅的閩南語句
- 輸出
  - 斷詞的全漢佮全羅的閩南語句

斷詞前	自稱一世人離袂開預報工作的吳德榮
	tsu7 tshing1 tsit8 si3 lang5 li5 be7 khui1 …
斷詞後	自稱 一世人 離袂開 預報 工作 的 吳德榮
	tsu7-tshing1 tsit8-si3-lang5 li5-be7-khui1 …

第二个問題 - 未知詞問題

- 拄著無看過的華語詞就會翻袂出來
  - 對齊模型、語言模型的單位攏是詞
  - 語料無全部的華語詞

訓練語料1	自稱 **一輩子** 離不開 預報 工作 的 吳德榮 ，
	tsu7-tshing1 **tsit8-si3-lang5** li5-be7-khui1 …
訓練語料2	開始 傷**愁** 了 ，
	khai1-si2 **siong1-tshiu5** ah4 ，
……	

試驗輸入	**一輩子** 吃 穿 都 不用 **憂愁** 了
翻譯結果	**tsit8-si3-lang5** tsiah8 tshing7 long2 m7-bian2 憂愁

第三个問題 - 語料整理

- 輸入
  - 無完整的閩南語句
- 輸出
  - 有全漢、全羅、斷詞的閩南語句

這幾工寒流閣再展威
tsit4 kui2 kang1 han5 liu5 koh4 tsai3…

這 幾工 寒流 閣再 展威
tsit4 kui2-kang1 han5-liu5 koh4-tsai3…

Koh m7知u7危險
Koh m7-tsai u7 gui5-hiam2

閣 毋-知 有 危-險
Koh m7-tsai u7 gui5-hiam2

第四个問題 - 語言分類

- 動機
  - 網路頂的語料，閩南語定定濫華語
- 輸入：一段語句
- 輸出：語句是閩南語，抑是華語

語句	語言
聽人講 khah 早有出現過『小蜜蜂』	閩南語
我 beh tńg 來種作！── 記 0312 Truku 反亞泥 • 還我土地運動	閩南語
有台灣味 ê 繪本──《我和我的腳踏車》.	閩南語
「 糟了 ，是工地火燒厝， 緊轉去打 火 ！ 」建設公司 的 工地主任 從手機接到消息，通話結束後就帶著那群混混先離開了。	華語
去越南胡志明市 4 工 / 越南胡志明市四日行 @Gio̍k-hōng	華語

第四節：研究方法佮實驗結果

- 閩南語斷詞
  - 實驗一：斷詞效果
- 未知詞問題
  - 實驗二：斷字佮斷詞的效果比較
- 語料整理
  - 實驗三：校對的效果
- 語言分類
  - 實驗四：語言分類效果
  - 實驗五：加入TGB的翻譯效果

閩南語斷詞 - 拄好長度斷詞方法

- 目的
  - 減少長詞優先的錯誤
  - 字平均分配到逐个詞
  - 詞數愈少愈好

長詞優先（後壁）	甚至 和 **國 小學生** 嘛 想 袂 開
答案	甚至 和 **國小 學生** 嘛 想 袂 開
長詞優先（頭前）	猶 掠做 **唱歌 仔 戲** 真 簡單
答案	猶 掠做 **唱 歌仔戲** 真簡單

- 方法
  - 用維特比（Viterbi）揣出成本上低的斷詞切法
  - 要求成本愈低愈好
    - 成本函數
      - n字詞成本$^1/_n$
  - 若一句話攏總m字，$j_1, j_2, \cdots, j_m$
  - 拄好長度斷詞(m)
    - *若*$[j_{i+1}, j_{i+2}, \cdots, j_m]$是一个詞
    - 斷詞位置 $= \underset{i}{\operatorname{argmin}}\{^1/_{m-i} + \text{拄好長度斷詞}(i)\}$

閩南語斷詞 - 拄好長度斷詞範例

- 成本範例

方法	斷詞	成本
長詞優先（後壁）	甚至 和 **國 小學生** 嘛 想 袂 開	…+1/1+1/3+…
答案	甚至 和 **國小 學生** 嘛 想 袂 開	…+1/2+1/2+…
長詞優先（頭前）	猶 掠做 **唱歌 仔 戲** 真 簡單	…+1/2+1/1+1/1+…
答案	猶 掠做 **唱 歌仔戲** 真簡單	…+1/1+1/3+…

- 長詞較好的例

答案	七月半 鴨仔 毋 知 死活
拄好長度	七 月半 鴨仔 毋知死 活
長詞優先	七 月半 鴨仔 毋 知 死活

實驗一 - 斷詞效果

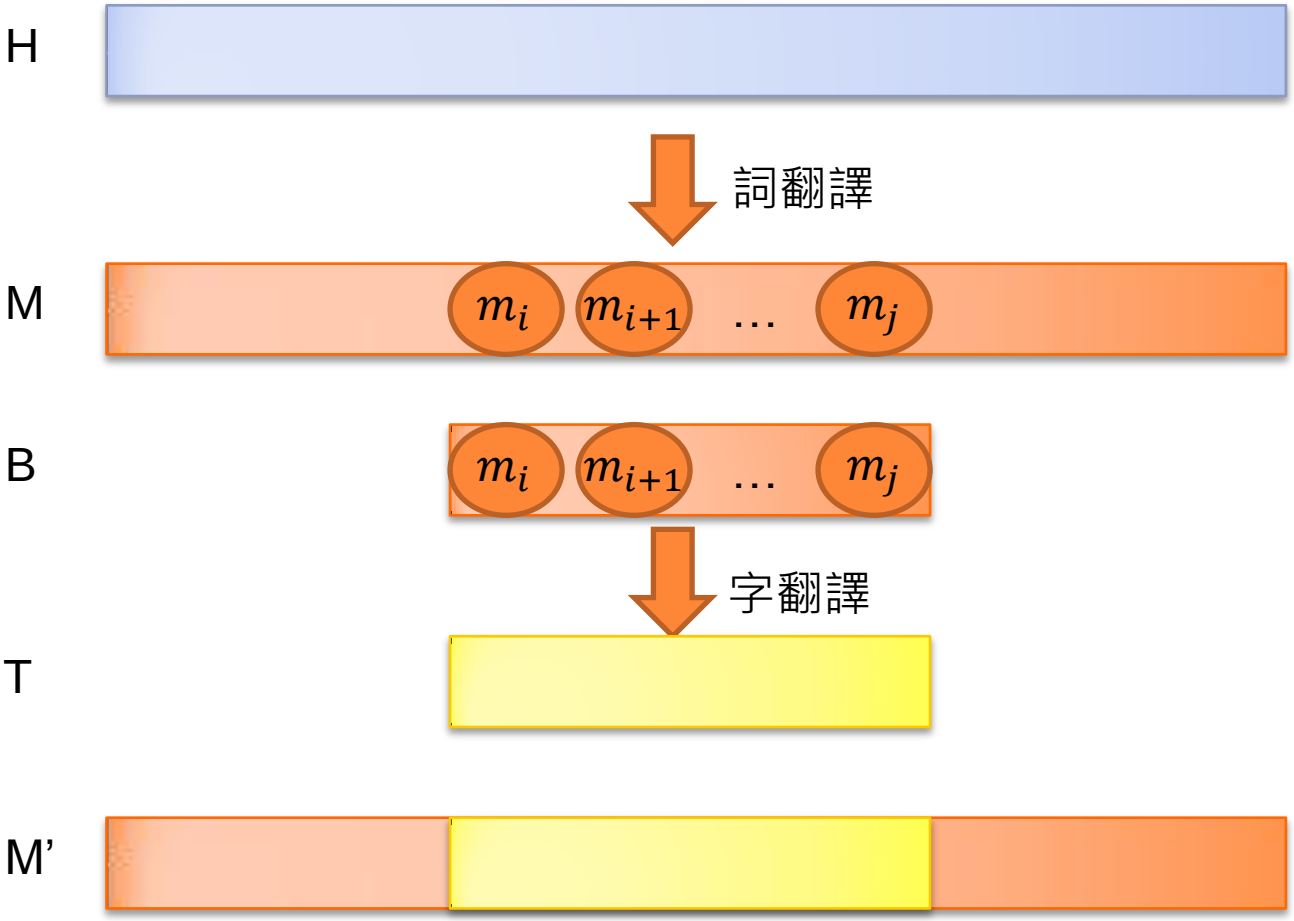
- 訓練語料
  - 教育部辭典條目35130詞
- 試驗語料
  - 教育部辭典例句8027句

斷詞方法	召回率	精確率	F -測量
拄好長度斷詞	91.1	85.1	88.0
長詞優先（對頭前）	91.0	84.9	87.9
長詞優先（對後壁）	91.1	85.0	88.0

未知詞問題 - 未知詞另外翻譯方法

- 方法
  1. 原本華語句$H = \{h_1, h_2, \ldots\}$
     - h_i代表一个華語詞
  2. 先用詞翻譯，得著$M = \{m_1, m_2, \ldots m_n\}$
     - m_i代表一个閩南語詞
  3. 若B $= \{m_i, m_{i+1}, \ldots m_j\}$ 攏是未知詞，m_{i-1}, m_{j+1} 是已知詞
  4. B提去字翻譯得著T
  5. 共M內的B換做T，得著M'
  6. 重做第3步，到無B存在為止

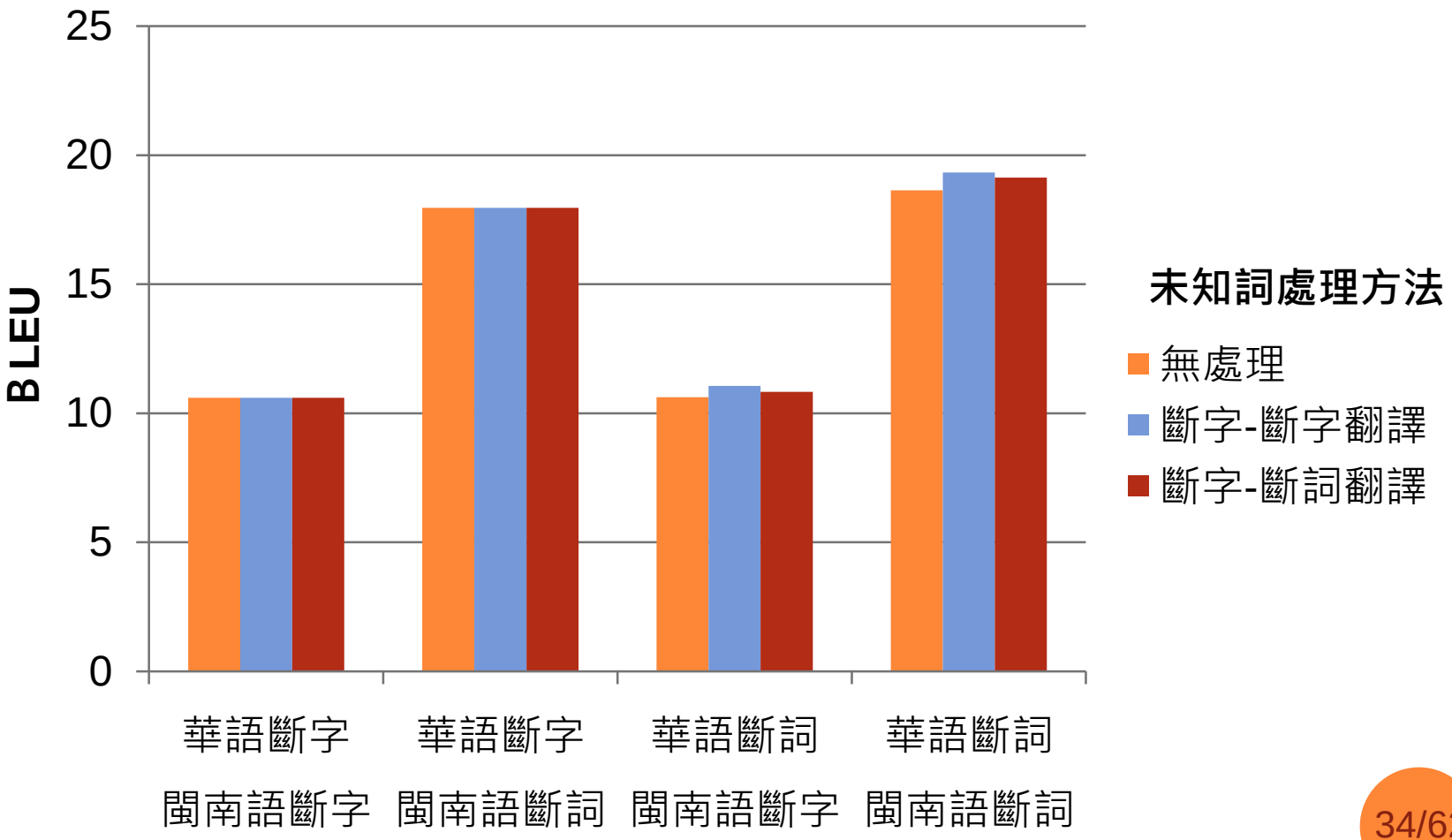
未知詞問題 - 未知詞另外翻譯流程



實驗二 - 斷字佮斷詞的效果比較環境

- 訓練語料
  - 教育部辭典條目35130詞
  - 教育部附錄句388句
  - 新聞平行語料64121句
  - 臺文典藏329476句
  - TGB平行語料35025句
- 試驗語料
  - 教育部辭典平行例句8027句
- 翻譯效果
  - BLEU分數以詞為單位

實驗二 - 斷字佮斷詞的效果比較實驗



整理語料 - 樣式一致

	全漢	全羅	斷詞	
教育部辭典	O	O	O	彼 个 查某 囡仔 真媠 。
				Hit e5 tsa-boo2 gin2-a2 tsin sui2 .
新聞語料庫	O	O	X*	嘛向望 老母 身體 勇起來
				ma7-ng3-bang7 lau7-bu2 sin1-the2 iong2-khi2-lai5
臺文典藏	X	O**	O	Koh m7知u7危險
				Koh m7-tsai u7 gui5-hiam2

嘛 向望 老母 身體 勇 起來
ma7 ng3-bang7 lau7-bu2 sin1-the2 iong2 khi2-lai5
閣 毋知 有 危險
Koh m7-tsai u7 gui5-hiam2

註*：新聞語料斷詞無規範
註**：臺文典藏少數無全羅

語料整理 - 語料簡寫

- 漢字佮音標簡寫做一逝
  - 若無仝，攏寫出來
  - 若仝款，寫一个

- 為著後壁投影片好解說

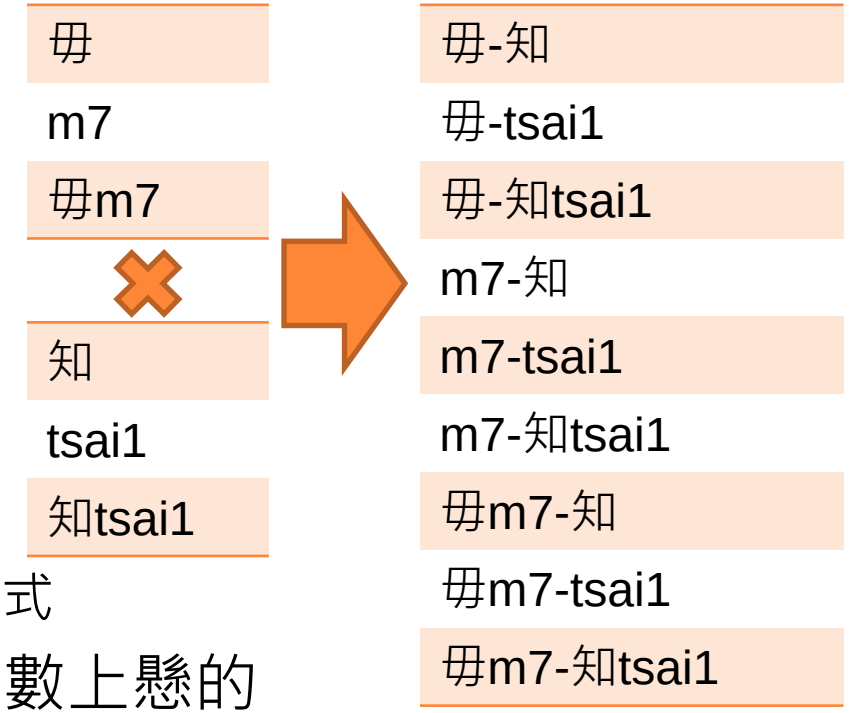
Koh4 m7知u7危險

Koh4 m7-tsai u7 gui5-hiam2

Koh4 m7-知tsai u7危gui5-險hiam2

語料整理 - 臺文典藏標漢字

- 問題
  - 無全部的字攏有漢字
    - 典藏的是漢羅佮全羅
  - 漢字用字佮教育部無仝
    - 字提掉
- 方法
  1. 仝款用拄好長度斷詞
     - 辭典有全漢、全羅、漢羅全部形式
  2. 確定斷點後，選語言模型分數上懸的
  3. 查著了後，照原本全羅斷詞



語料整理 - 整理一開始

新聞語料

臺文典藏

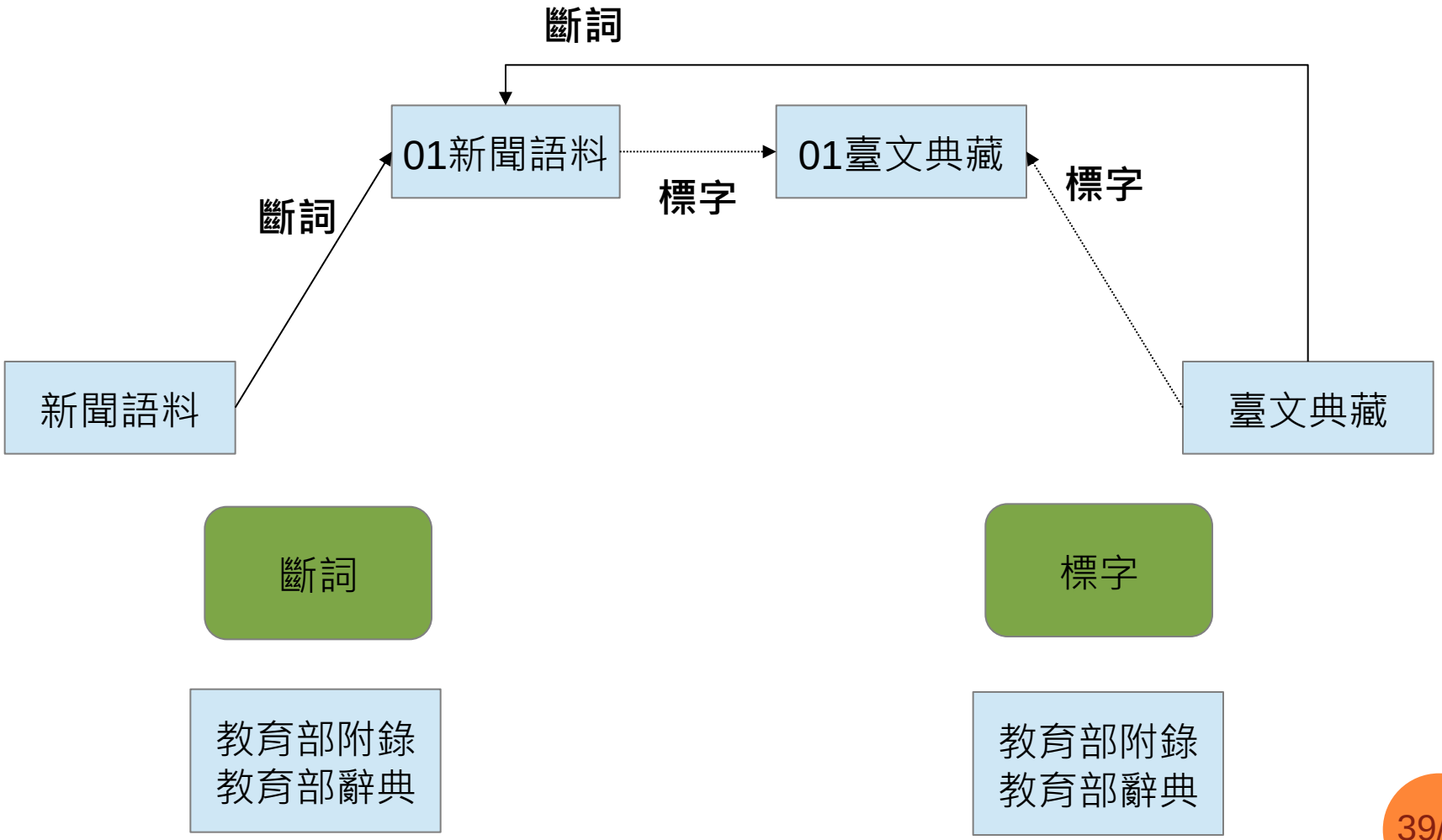
斷詞

標字

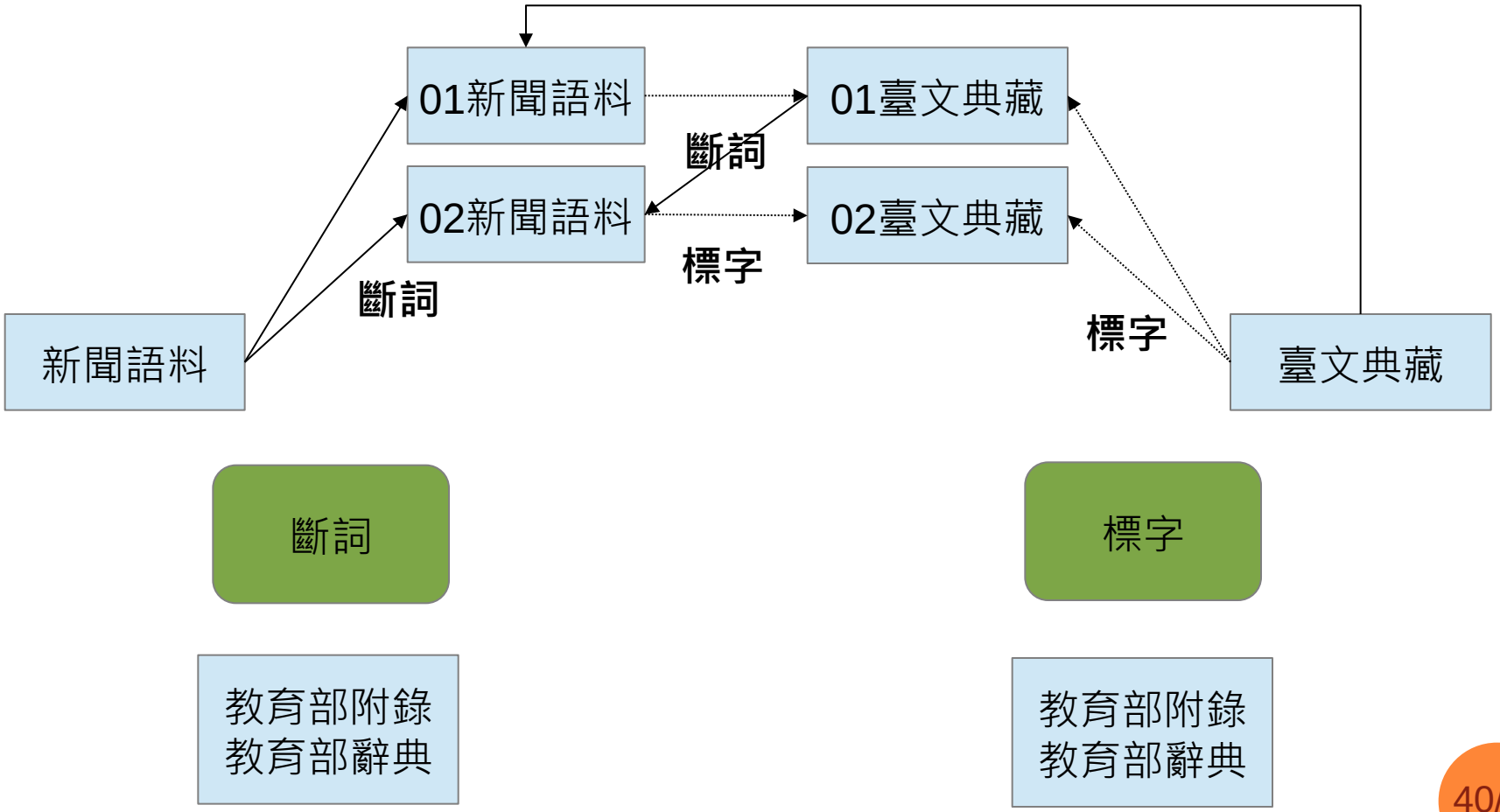
教育部附錄
教育部辭典

教育部附錄
教育部辭典

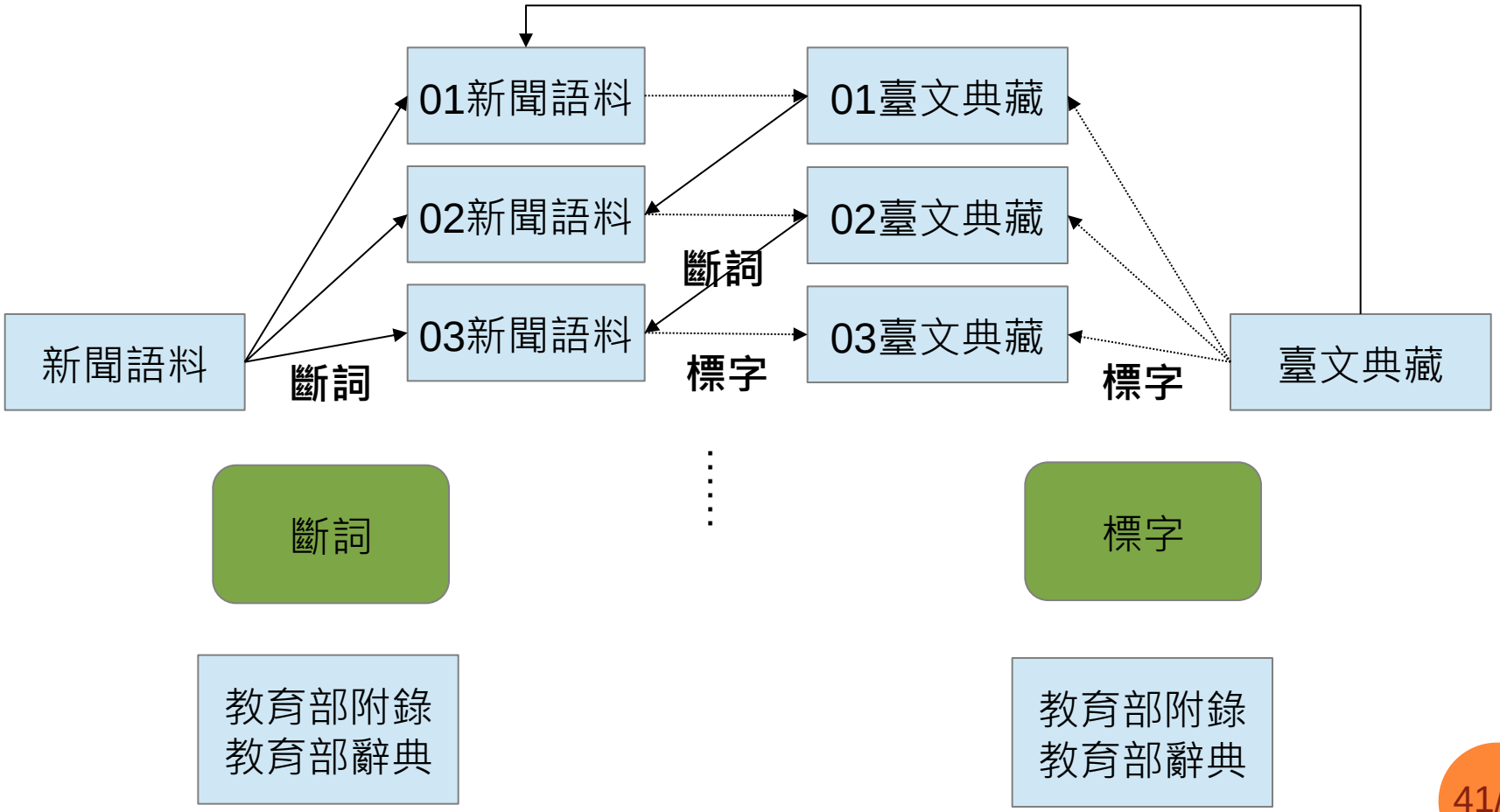
語料整理 - 整理第一擺


斷詞
01新聞語料
01臺文典藏
標字
斷詞
標字
新聞語料
臺文典藏
斷詞
標字
教育部附錄
教育部辭典
教育部附錄
教育部辭典

語料整理 - 整理第二擺

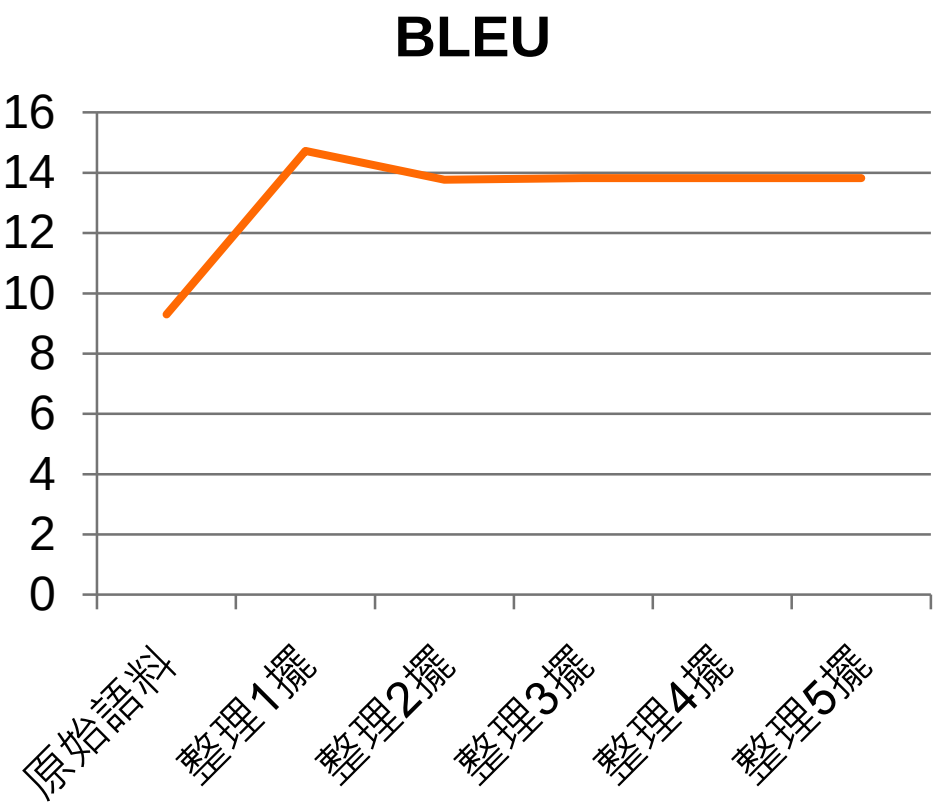

01新聞語料
02新聞語料
01臺文典藏
02臺文典藏
斷詞
標字
斷詞
標字
新聞語料
臺文典藏
斷詞
標字
教育部附錄
教育部辭典
教育部附錄
教育部辭典

整理語料 - 整理第三擺


01新聞語料
02新聞語料
03新聞語料
01臺文典藏
02臺文典藏
03臺文典藏
新聞語料
臺文典藏
斷詞
標字
斷詞
標字
斷詞
標字
教育部附錄
教育部辭典
教育部附錄
教育部辭典

實驗三 - 校對的效果

- 訓練語料
  - 教育部辭典條目35130詞
  - 教育部附錄388句
  - 新聞平行語料64121句
  - 臺文典藏329476句
- 試驗語料
  - 教育部辭典例句8027句
- 翻譯效果
  - BLEU分數以詞為單位



整理幾擺	原始	1	2	3	4	5
BLEU分數	9.30	14.72	13.77	13.82	13.82	13.82

實驗三 - 校對的語料

- 錯誤的例

原始語料	指tsi2 用ing7 二ji7-十tsap8-三sann1 **e5** 字ji7-母bu2 tiann7-tiann7
整理1擺	指tsi2 用ing7 二ji7-十tsap8-三sann1 **个e5** 字ji7-母bu2 定tiann7-定tiann7
整理2擺	指tsi2 用ing7 二ji7-十tsap8-三sann1 **的e5** 字ji7-母bu2 定tiann7-定tiann7

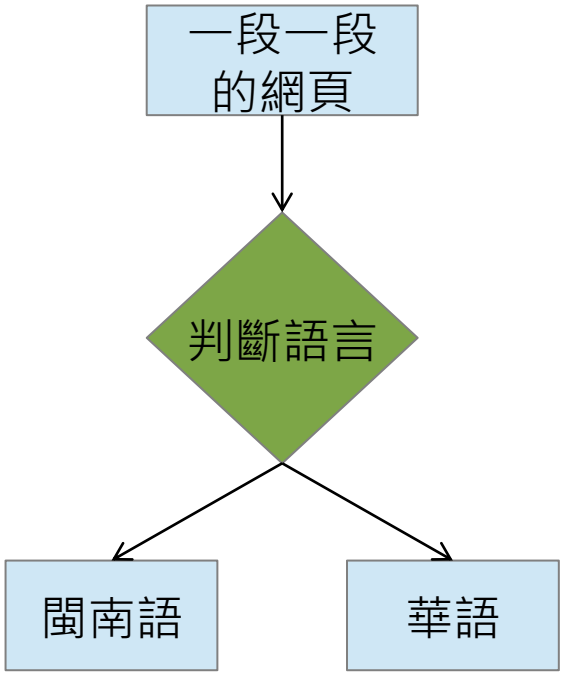
- 頭一擺整理時，辭典內底無「二十三」這个詞，所以揀出「三 个」
- 第二擺整理時，「二十三 个」無出現過

原始語料	佇ti7 已i2-經king1 開khui1-出tshut4 的e5 選suan2-票phio3 中tiong1
整理1擺	佇ti7 已i2-經king1 **開khui1-出tshut4 的e5** 選suan2-票phio3 中tiong1
整理2擺	佇ti7 已i2-經king1 **開khui1 出tshut4-的e5** 選suan2-票phio3 中tiong1

- 語料內底有「是 對 海關 出的」

分類語言 - 問題

- 問題
  - 以早判斷語言的研究
    - 對象是拼音文字為主
      - 用字元算語言模型分數
    - 無適合用佇分閩南語華語
      - 閩南語華語有誠濟共同詞
- 解決方法
  - 利用斷詞佮定用詞做特徵
    - 斷詞的詞數資訊
    - 定用詞愛提掉共同詞
      - 後壁號做「特徵詞」



分類語言 - 特徵詞介紹

- 特徵詞
  - 無共同詞的定用詞

- 方法
  1. 閩南語佮華語分別選n个上定用的詞
  2. 揣閩南語頭前m个無出現佇華語n个的定用詞
     - 就揣出閩南語m特徵詞
  3. 揣華語頭前m个無出現佇閩南語n个的定用詞
     - 就揣出華語m特徵詞

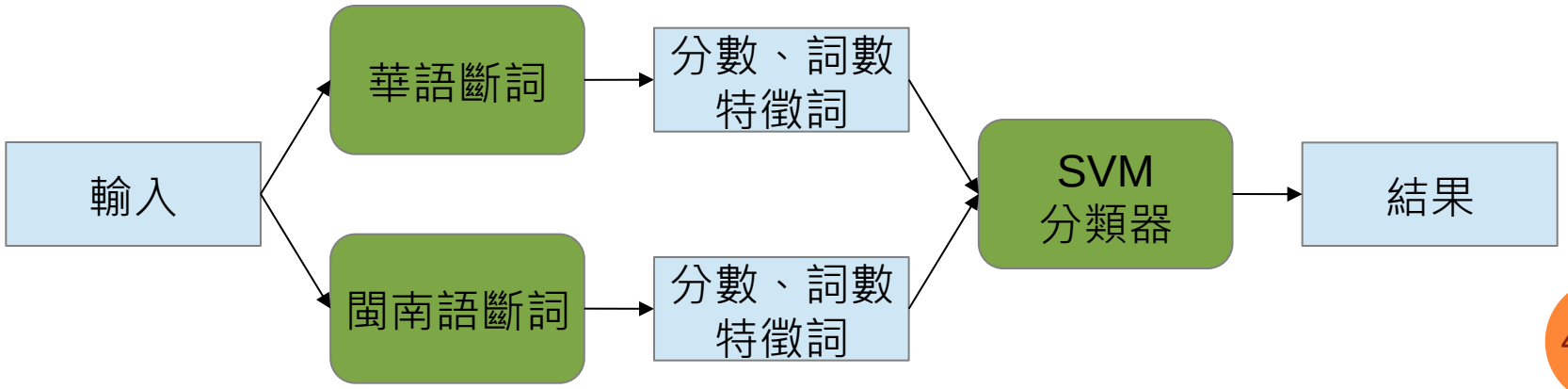
分類語言 - 特徵詞範例

- 閩南語統計來源
  - 教育部例句附錄句、新聞語料庫、臺文典藏
- 華語統計來源
  - 中研院1000萬字平衡語料庫
- 定用詞數量n=7000，特徵詞數量m=3000

閩南語 定用詞	的 e5	伊 i1	有 u7	是 si7	我 gua2	人 lang5	無 bo5	講 kong2	伫 ti7	…
華語 定用詞	的	是	在	一	有	了	不	我	個	…
閩南語 特徵詞	伫 ti7	个 e5	閣 koh4	攏 long2	佮 kap4	口 in1	咧 teh4	咱 lan2	彼 hit4	…
華語 特徵詞	我們	很	她	沒有	或	他們	更	則	把	…

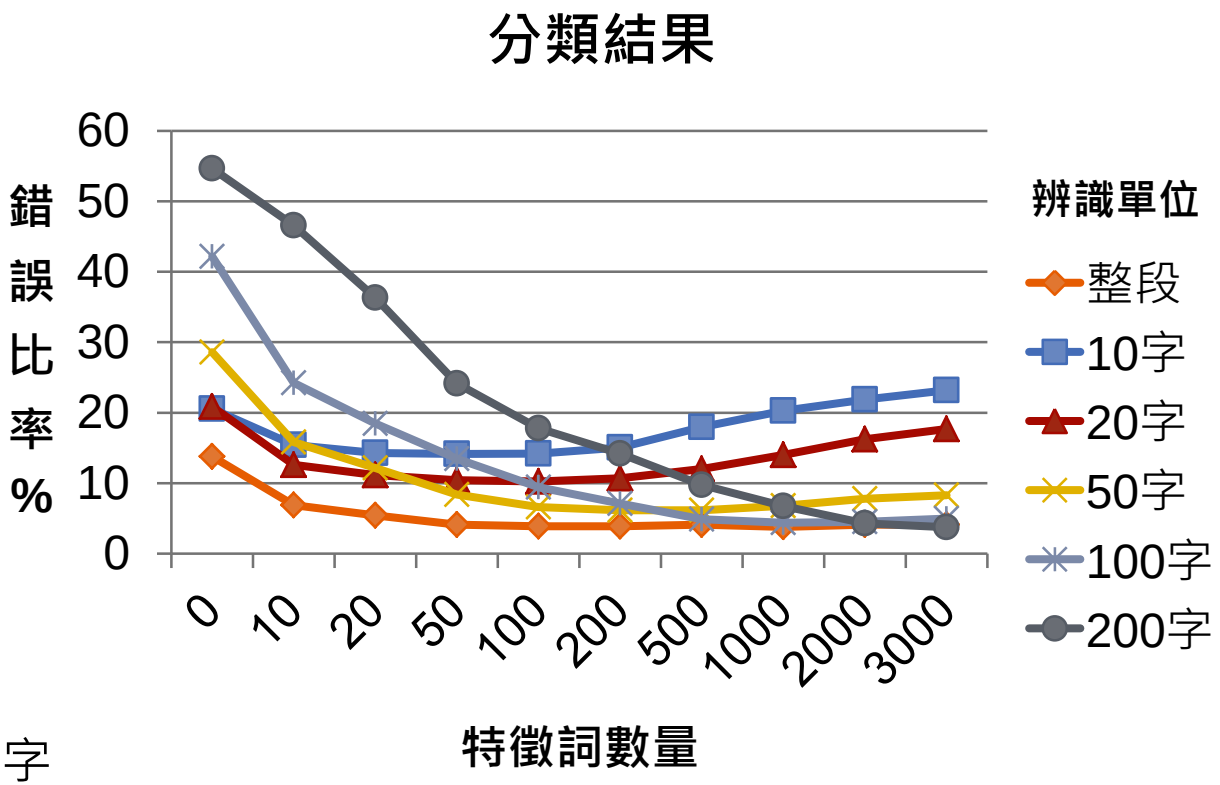
分類語言 - 參數

- 特徵
  - 語言模型分數
  - 斷詞詞數
  - 1~k字詞分別數量
  - m特徵詞
    - m若傷大，會影響著速度



實驗四 - 語言分類效果

- TGB通訊語料
- 實驗參數
  - 字詞數量
    - k=4
  - 定用詞數量
    - n=7000
  - 特徵詞數量
    - m=0~3000
  - 辨識單位
    - 段
      - 華語平均74字
      - 閩南語平均51字
    - n字
      - 連紲選k句
      - 總字數超過n



	總合	華語	閩南語
訓練	1000篇文章	8519段	9368段
試驗	179篇文章	2397段	1344段

實驗五 - 加TGB語料的翻譯效果

- 訓練語料
  - 教育部辭典條目35130詞
  - 教育部附錄句388句
  - 新聞平行語料64121句
  - 臺文典藏329476句
  - **TGB平行語料35025句**
- 試驗語料
  - 教育部辭典平行例句8027句
- 翻譯效果
  - BLEU分數以詞為單位

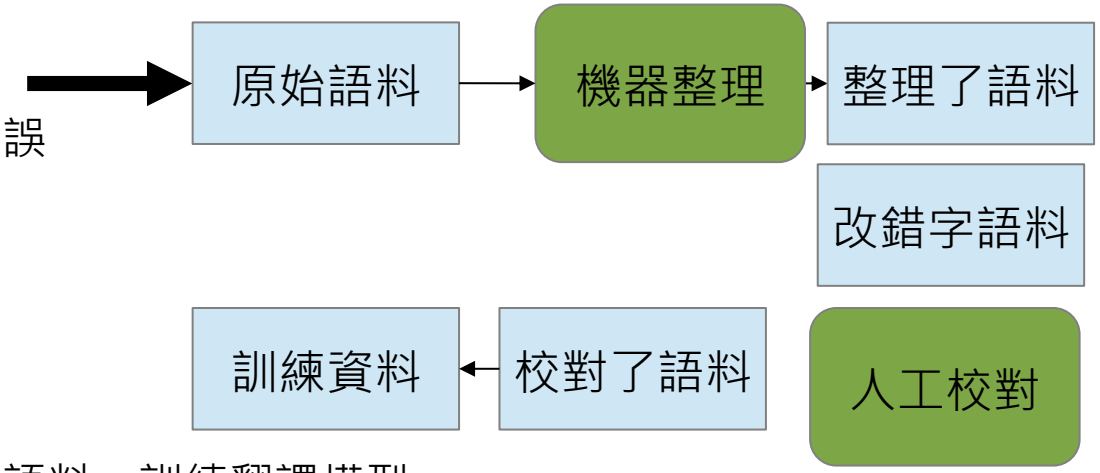
	加TGB語料前	加TGB語料後
平行語料句數	64121句	99146句
BLEU分數	13.82	19.33

第五節：結論佮未來發展

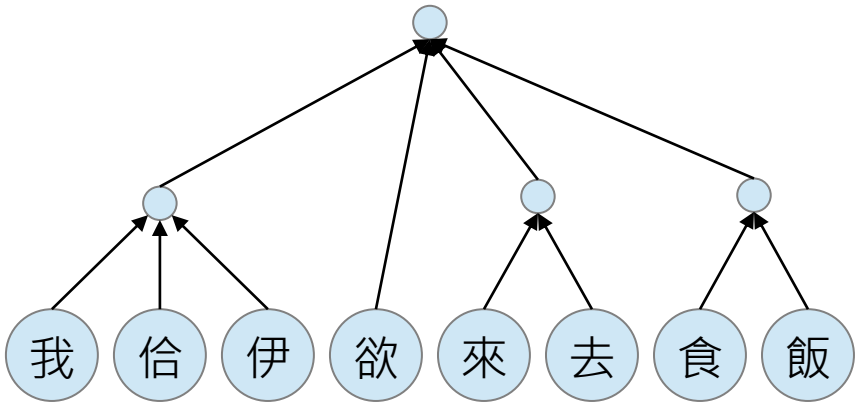
- 結論
  - 拄好長度斷詞佮長詞優先差無濟
  - 用斷字翻譯有改善未知詞問題
  - 分類閩南語佮華語各50个特徵詞就夠用
  - 整理資訊無完整的語料庫，對翻譯有幫助
  - 語料的數量影響效果誠濟
- 未來發展
  - 自動校對語料
  - 加強斷詞
  - 應用佇字幕辨識

未來發展—加強翻譯

- 自動校對語料
  - 目的
    - 翻譯的訓練語料有誠濟錯誤
    - 語料需要人工校對
    - 減人工負擔
  - 問題
    - 語料改錯字
  - 方法
    - 用「原始語料」佮「校對語料」訓練翻譯模型



- 加強斷詞
  - 目的
    - 斷詞效果影響著翻譯效果
  - 方法
    - 詞性斷詞
    - 剖析器
      - 我佮伊欲來去食飯



未來發展—應用

- 字幕辨識
  - 目的
    - 聲音語料大部份母語發音、配華語字幕
      - 電視劇、廣播
    - 補上母語字幕，學母語用字
  - 問題
    - 輸入母語聲音、華語字幕
    - 輸出母語字幕

第六節：參考文獻

- Ma, Wei-Yun and Keh-Jiann Chen, 2003, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff", Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pp168-171.
- Peter F. Brown , Vincent J. Della Pietra , Stephen A. Della Pietra , Robert L. Mercer, The mathematics of statistical machine translation: parameter estimation, Computational Linguistics, v.19 n.2, June 1993
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

實際翻譯例

食安問題連環爆，立法院長王金平下午主持《食品安全衛生管理法》協商仍未獲共識

食安問題連環爆，立法院長王金平下晡主持《食品安全衛生管理法》協商猶未獲共識

昨晚在粉絲團貼篇北海豬油製品流入大潤發等賣場的分享文

昨暗佇粉絲團貼篇北海豬油製品流入大純發等賣場的分享文

究竟傳統是甚麼?有沒有可以讓我們依循的道路?

究竟傳統是按怎？有無通予咱照的道路？

如果傳統的重要族群活動，看不到文化，體會不到精神，

若是傳統的重要的族群活動，看袂著文化，體會毋著精神，

多謝逐家

附錄

- 對齊模型介紹
- 語言模型介紹

對齊模型介紹

- 目的
  - 揣出華語佮閩南語的詞機率對應
  - 功能親像雙語辭典
    - 對應資訊是對平行語料揣出來的
- 做法
  1. 平行語料一句一句處理
  2. 記錄華語全部的詞佮閩南語全部的詞對應幾擺
  3. 對應較濟擺的對應組合就是翻譯選項之一

對齊模型範例

- 輸入語料
  1. 他 打 我/伊 共 我 拍
     - 他-伊/他-共/他-我/他-拍
     - 打-伊/打-共/打-我/打-拍
     - 我-伊/我-共/我-我/我-拍
  2. 打 鼓 很 好玩/拍 鼓 誠 趣味
     - 打-拍/打-鼓/打- 誠/打-趣味
     - .....
- 「打」的對應結果
  - 語料是整理過的

華-閩南對應	對應機率	華-閩南對應	對應機率
打-伊	1/8	打-鼓	1/8
打-共	1/8	打-誠	1/8
打-我	1/8	打-趣味	1/8
打-拍	2/8		

對齊模型種類無仝範例

- 輸入語料
  1. 他 打 我/伊 共 我 phah
     - 他-伊/他-共/他-我/他-phah
     - 打-伊/打-共/打-我/打-phah
     - 我-伊/我-共/我-我/我-phah
  2. 打 鼓 很 好玩/拍 鼓 誠 趣味
     - 打-拍/打-鼓/打- 誠/打-趣味
     - .....
- 「打」的對應結果
  - 猶未整理的語料
  - 後壁愛處理的問題

華-閩南對應	對應機率	華-閩南對應	對應機率
打-伊	1/8	打-鼓	1/8
打-共	1/8	打-誠	1/8
打-我	1/8	打-趣味	1/8
打-拍	1/8	打-phah	1/8

語言模型介紹

- 目的
  - 判斷語句合理性
  - 語句分做細句，合理性就是逐个細句機率相乘
- 訓練方法
  - 先決定一个數字n，代表一擺看n个連緌的詞
  - 共一句k个詞的句，產生k-n+1組的n連緌詞
  - 統計全部連緌詞的數量
- 判斷方法
  - 仝款共試驗語句轉做n連緌詞
  - 對逐个連緌詞，用頭前n-1个詞，算第n个詞出現的條件機率
    - $\mathrm{p}(詞_n|詞_1, 詞_2,\ldots, 詞_{n-1})$
  - 語句的合理性是全部n連緌詞的機率乘起來
    - $\mathrm{p}(一句話) = p(詞_1, 詞_2, \ldots, 詞_k)$
      $\sim \prod_{i=1}^{k-n+1} p(詞_{i+k-1}|詞_i, 詞_{i+1},\ldots, 詞_{i+n-1})$

語言模型範例 - 訓練

- 假設n=2，考慮二連緌詞
- 輸入語料
  1. 伊 共 我 **拍**
  2. **拍** 鼓 誠 趣味
  3. 我 **敲** 電話 予 伊

連緌詞	對應機率	連緌詞	對應機率
p([句尾]|拍)	1/2	p(拍|我)	1/2
p(鼓|拍)	1/2	p(敲|我)	1/2
p(電話|拍)	10^{-10*}	p(我|[句頭])	1/3
p([句尾]|鼓)	10^{-10*}	p([句尾]|電話)	10^{-10*}

*註：無出現過的機率，有專門的算法予機率加起來是1

語言模型範例 - 使用

- 假設n=2，考慮二連綴詞
- 語句機率
  - 我 拍 鼓
    - $p(\text{我}|[\text{句頭}]) \times p(\text{拍}|\text{我}) \times p(\text{鼓}|\text{拍}) \times p([\text{句尾}]|\text{鼓})$
  - 我 拍 電話
    - $p(\text{我}|[\text{句頭}]) \times p(\text{拍}|\text{我}) \times p(\text{電話}|\text{拍}) \times p([\text{句尾}]|\text{電話})$

連綴詞	對應機率	連綴詞	對應機率
$p([\text{句尾}]\|\text{拍})$	1/2	$p(\text{拍}\|\text{我})$	1/2
$p(\text{鼓}\|\text{拍})$	1/2	$p(\text{敲}\|\text{我})$	1/2
$p(\text{電話}\|\text{拍})$	10^{-10*}	$p(\text{我}\|[\text{句頭}])$	1/3
$p([\text{句尾}]\|\text{鼓})$	10^{-10*}	$p([\text{句尾}]\|\text{電話})$	10^{-10*}

*註：無出現過的機率，有專門的算法予機率加起來是1