

漢語間統計式機器翻譯語料處理 用臺灣閩南語示範

Corpus Preprocessing for Statistical Machine
Translation between the Chinese Languages
Using Taiwan Southern Min as Examples

103/10/30

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
- 第二節：相關文獻與背景智識
- 第三節：研究方法
- 第四節：結論與未來發展
- 第五節：參考文獻
- 附錄

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
 - 予欲學的人參考
 - 配合語音合成
- 針對漢語之間的翻譯
 - 主要處理翻譯語料
 - 無修改翻譯演算法
 - 用閩南語示範，結果嘛會使用佇客話

第二節：相關文獻背景智識

- 語料狀況
- 語料庫介紹
- 語料樣式
- 翻譯模型
- 評分方式

閩南語語料種類

種類	範例	語料庫	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	漢字恰拼音擁有
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

- 逐個語料庫攏無仝
 - 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
- 內部有規範
 - 資料攏有全漢佮全羅
 - 音標有斷詞
 - 有腔口標示
- 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	彼个查某囡仔真嬌。
----	-----------

全羅	Hit ê tsa-bóo gín-á tsin suí.
----	-------------------------------

華語	那個女孩子很漂亮。
----	-----------

全漢	我欲去買雞卵。
----	---------

混合腔全羅	Guá beh khì bé ke-nṅg.
-------	------------------------

偏泉腔全羅	Guá beh khì bé kue-nṅg.
-------	-------------------------

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
- 有華語恰對應的閩南語全漢全羅
- 翻譯時，罕得調整語詞先後
- 現代用語、古早用語攏有
- 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－數位典藏

- 全名「台語文數位典藏」
- 國家臺灣文學館收集1885～2006年的語料
- 漢羅佮全羅對照
 - 原本干焦一種，臺文館後來倩人拍字
 - 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

- 形式真濟款

- 華語
- 閩南語
- 華語閩南語平行
- 華語閩南語濫做伙

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？

濫做伙範例

『聽說妳最近遇到什麼問題，是不是？怎麼了？』好性地 ê QA 繼續問--落-去。

- 1999年到這馬

- 實驗的資料到2014年6月12日，1179篇文章

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
 - 數位典藏
 - 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
數位典藏	229/135	458/365	703/390

語料樣式

○ 語料樣式

● 斷字

- 無詞的資訊

● 斷詞

- 華語

- 中研院中文斷詞系統 (CKIP)

- [1] Ma, Wei-Yun, 2003

- 閩南語

- 長詞優先斷詞

○ 後壁會比較這兩個對翻譯的影響

- 攏會提來試看覓

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

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
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

長詞優先斷詞範例

○ 範例 - 蔡崇名細漢時生活困苦

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.

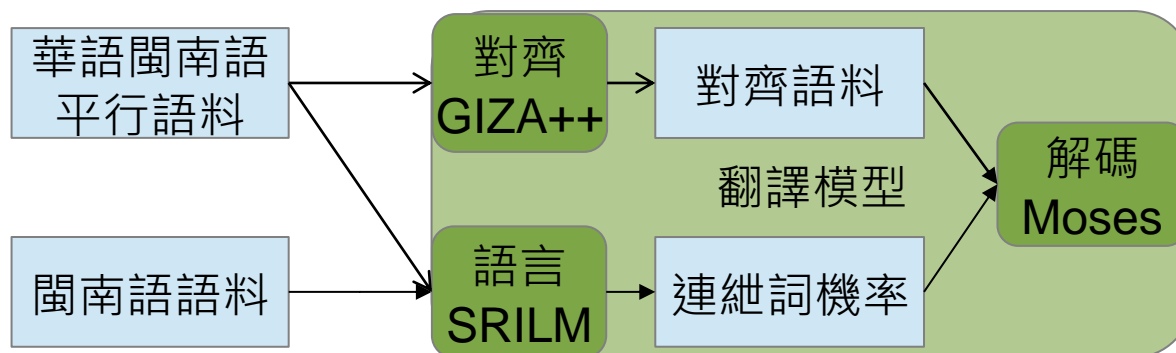
斷詞評分方式

- 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
 - 華語佾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
 - 用對齊模型佾語言模型翻譯
 - Philipp Koehn et al. 2007.



對齊模型介紹

○ 目的

- 揣出華語佾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. 記錄華語全部的詞佾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，合理性就是逐個細句機率相乘

○ 訓練方法

- 先決定一個數字 n ，代表一擺看 n 個連繼的詞
- 共一句 k 個詞的句，產生 $k-n+1$ 組的 n 連繼詞
- 統計全部連繼詞的數量

○ 判斷方法

- 全款共試驗語句轉做 n 連繼詞
- 對逐個連繼詞，用頭前 $n-1$ 個詞，算第 n 個詞出現的條件機率
 - $p(\text{詞}_n | \text{詞}_1, \text{詞}_2, \dots, \text{詞}_{n-1})$
- 語句的合理性是全部 n 連繼詞的機率乘起來
 - $p(\text{一句話}) = p(\text{詞}_1, \text{詞}_2, \dots, \text{詞}_k)$
 - $\sim \prod_{i=1}^{k-n+1} p(\text{詞}_{i+n-1} | \text{詞}_i, \text{詞}_{i+1}, \dots, \text{詞}_{i+n-1})$

語言模型範例 - 訓練

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

連繼詞	對應機率	連繼詞	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

語言模型範例 - 使用

- 假設 $n=2$ ，考慮二連繼詞
- 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
 - $p(\text{我}|\text{[句頭]}) \times p(\text{拍}|\text{我}) \times p(\text{電話}|\text{拍}) \times p(\text{[句尾]}|\text{電話})$

連繼詞	對應機率	連繼詞	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

BLEU評分

- Bilingual Evaluation Understudy
- 以連繼詞n-grams為單位
- $BLEU =$

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

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

連繼詞 _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

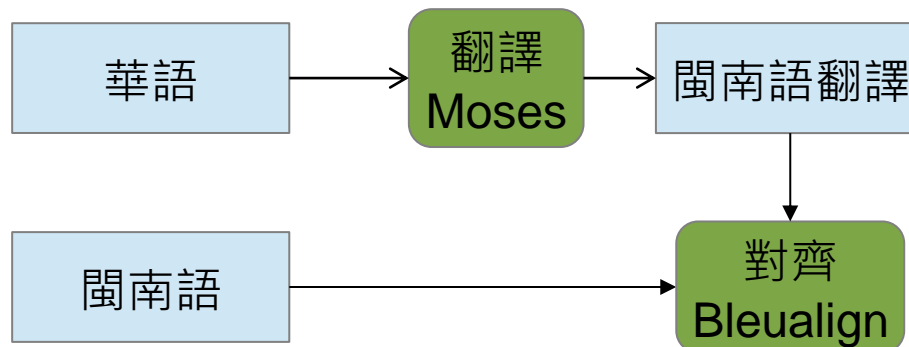
語料處理

○ 語言分類

- Cavnar and Trenkle, 1994
 - 用語言模型算分數
 - 以字元為單位

○ 語料對齊

- Sennrich and Volk, 2010
 - 用翻譯模型對齊語料



貢獻

- 比較漢語語料樣式對翻譯的影響
- 提出一个整理漢語語料的方法
- 分類兩種漢語的方法

第三節：研究方法

○ 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

- 第一個問題，按怎斷詞（閩南語斷詞）
- 第二個問題，有的詞翻袂出來（未知詞問題）

○ 語料愈濟愈好，所以加語料

- 第三個問題，按怎加入資料無完整的語料庫（整理語料）
- 第四個問題，網路語料需要分華語佮閩南語（分類語言）

第一个問題 - 閩南語斷詞

- 輸入
 - 全漢伶全羅的閩南語句
- 輸出
 - 斷詞的全漢伶全羅的閩南語句

第二個問題 - 未知詞問題

○ 原因

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

○ 輸入

- 華語句

○ 輸出

- 閩南語句

訓練語料1 動員了 一百五十位 志工，

liok8-siok8 siu1-tioh8 lak8-khin1 e5 soo2-tit4 kiau2-sue3-tuann1。

訓練語料2 陸續 增加 好幾家 店面。

liok8-siok8 tsing1-ka1 kui2-na7-king1 tiam3-bin7。

訓練語料3 台南 空軍基地 要在 十日 開放 參觀；

tong7-uan5-liau2 tsit8-pah4-goo7-tsap8-ui7 tsi3-kang1。

試驗輸入 陸續 開放 一百五十項 的 規費，

翻譯結果 liok8-siok8 khai1-hong3 一百五十項 e5 規費

第三个問題 - 整理語料

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

第四個問題 - 分類語言

- 輸入
 - 一段語句
- 輸出
 - 是閩南語，抑是華語

閩南語斷詞 - 拄好長度斷詞方法

○ 目的

- 希望會當閃避長詞優先的缺點
- 字平均分配到逐個詞
- 詞數愈少愈好

○ 方法

- 要求成本愈低愈好
- 對無全長度的詞分數無全
 - 一字詞成本1
 - 兩字詞成本 $1/2$
 - 三字詞成本 $1/3$
 - ...
 - n 字詞成本 $1/n$

長詞優先 (後壁)

答案

甚至和 國 小學生 嘛 想 袂 開

甚至和 國小 學生 嘛 想 袂 開

長詞優先 (頭前)

答案

猶 掠 做 唱歌 仔 戲 真 簡單

猶 掠 做 唱 歌仔戲 真簡單

閩南語斷詞 - 拄好長度斷詞範例

○ 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+...
拄好長度斷詞	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+...
拄好長度斷詞	猶 掠做 唱 歌仔戲 真簡單	...+1/1+1/3+...

○ 缺點

全羅	斷詞結果
hoo7 i1 tsut4-khi3 sng2	予伊出去耍
hoo7-i1 tsut4-khi3 sng2	雨衣出去耍

未知詞問題 - 未知詞另外翻譯

○ 方法

1. 原本華語句 $H = \{h_1, h_2, \dots\}$
 - h_i 代表一个華語詞
2. 先用斷詞翻譯，得著 $M = \{m_1, m_2, \dots, m_n\}$
 - m_i 代表一个閩南語詞
3. 若 $B = \{m_i, m_{i+1}, \dots, m_j\}$ 攏是未知詞， m_{i-1}, m_{j+1} 是已知詞
4. B 提去斷字翻譯得著 T
5. 共 M 內的 B 換做 T
6. 重做第3步，到無 B 存在為止

整理語料 - 語料無一致

- 教育部辭典
 - 斷詞
 - 全漢佮全羅
- 新聞語料庫
 - 無規範斷詞
 - 全漢佮全羅
- 數位典藏
 - 斷詞
 - 漢羅佮全羅

教育部辭典	彼个查某囡仔真嬌。
	Hit ê tsa-bóo gín-á tsin suí.
新聞語料庫	這幾工 寒流 閣再 展威
	tsit4-kui2-kang1 han5-liu5 koh4-tsai3 tian2-ui1
數位典藏	Koh m7知u7危險..... ,
	Koh m7-tsai u7 gui5-hiam2.....,

整理語料 - 數位典藏標漢字

問題

- 無全部的字擁有漢字
 - 典藏的是漢羅佮全羅
- 漢字用字佮教育部無仝
 - 字提掉

方法

- 仝款用拄好長度斷詞
- 斷詞時查全漢、全羅、漢羅全部形式
- 確定斷點後，選語言模型分數上懸的
- 查著了後，照原本全羅斷詞



整理語料 - 整理一開始

新聞語料

斷詞

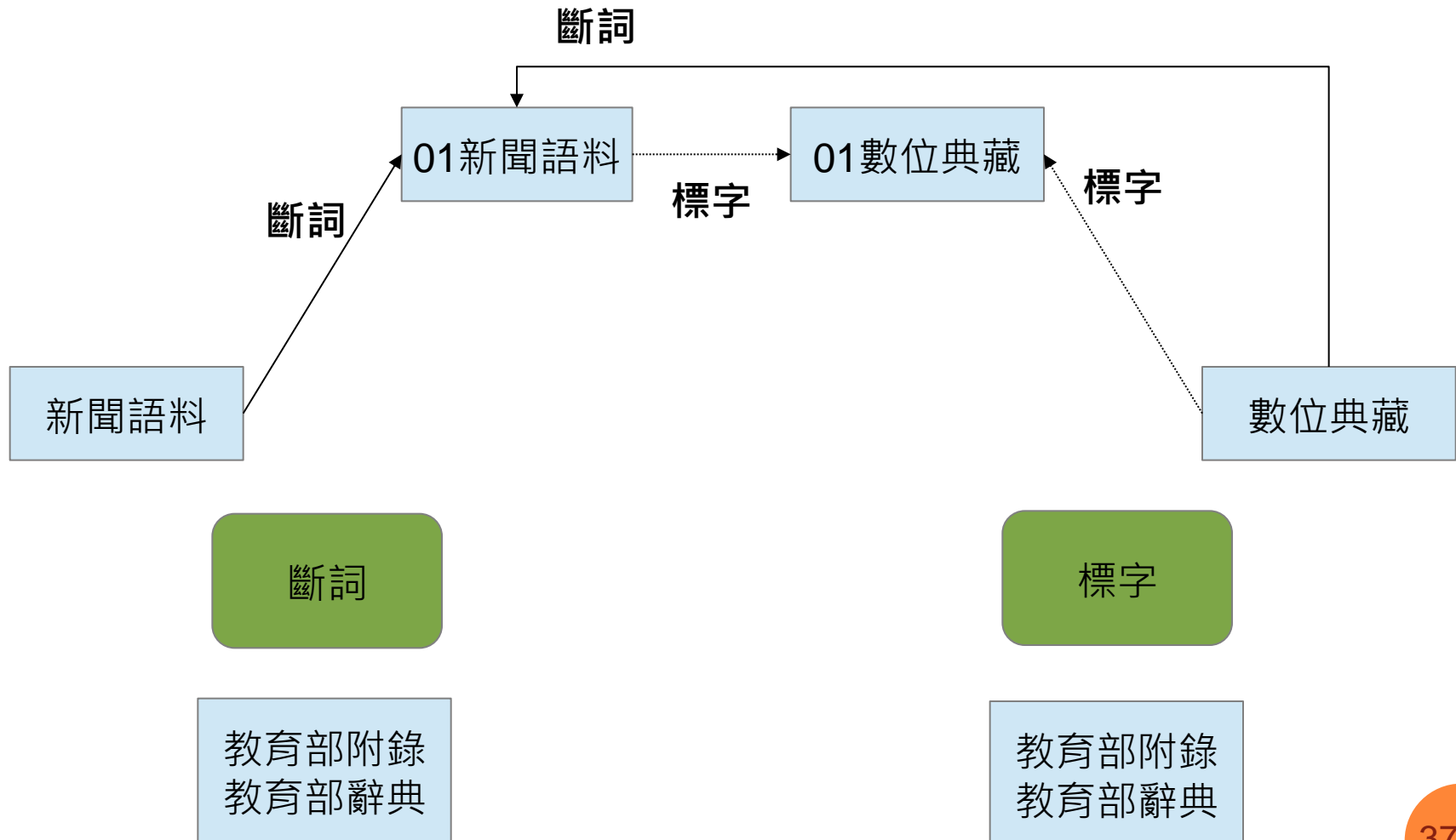
教育部附錄
教育部辭典

數位典藏

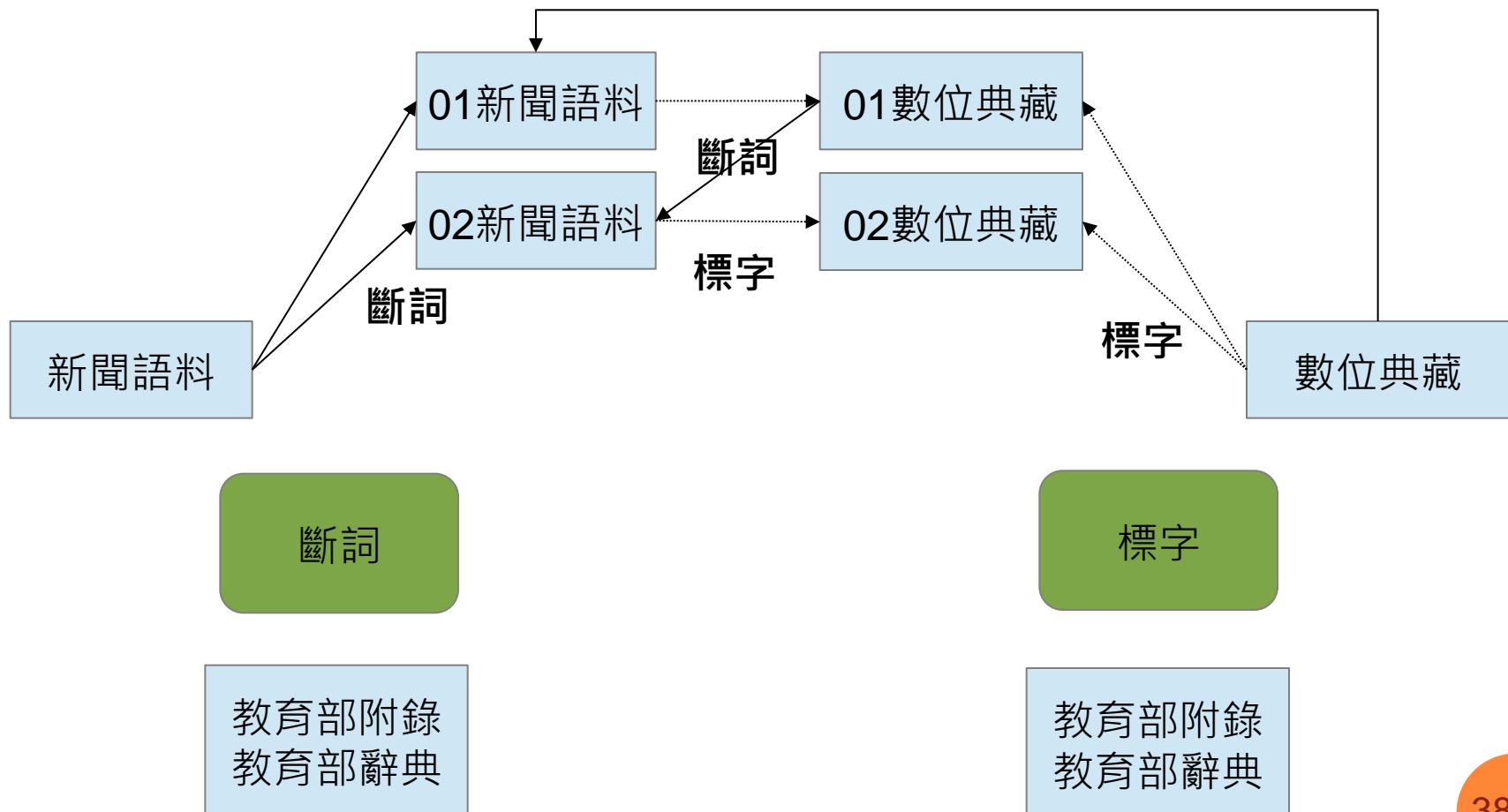
標字

教育部附錄
教育部辭典

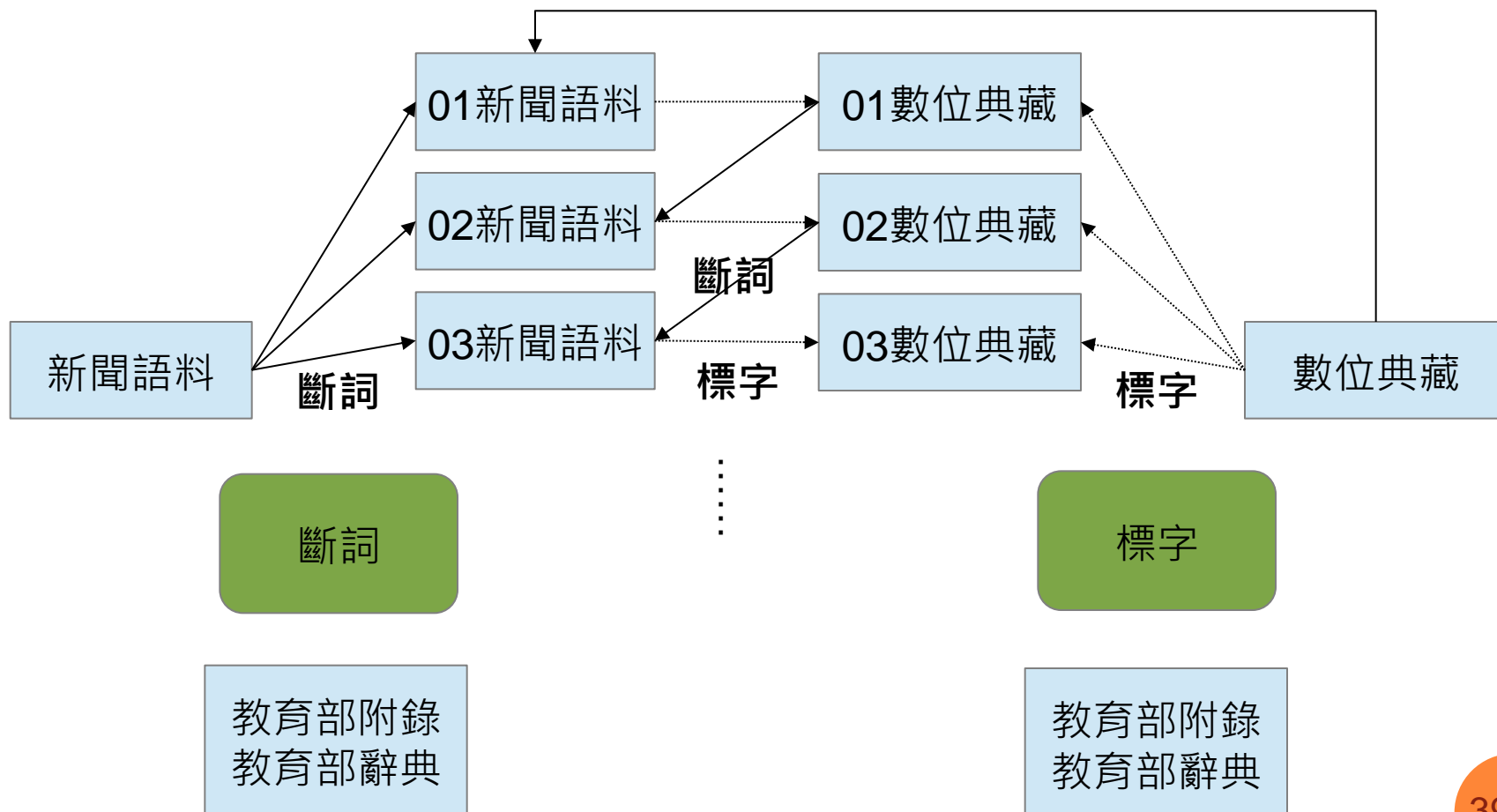
整理語料 - 整理第一擺



整理語料 - 整理第二擺



整理語料 - 整理第三擺



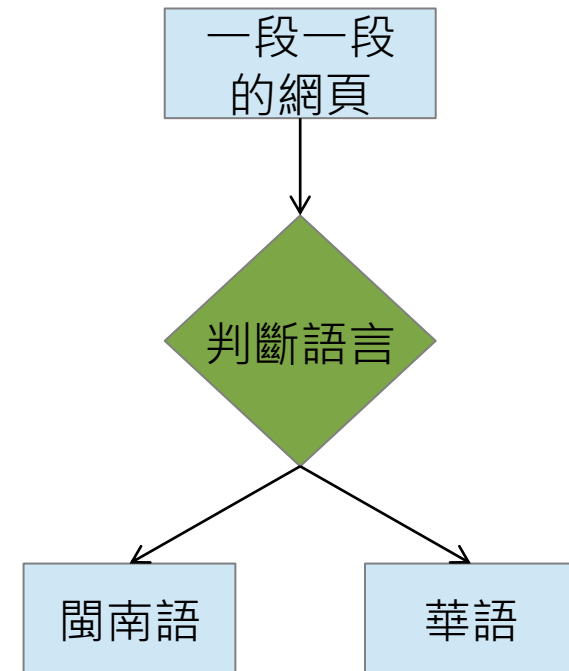
分類語言 - 方法

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

- 利用斷詞恰定用詞做特徵
 - 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

- 閩南語俚華語分別選 n 个上定用的詞
- 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
- 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、數位典藏

○ 華語統計來源

- 中研院1000萬字平衡語料庫

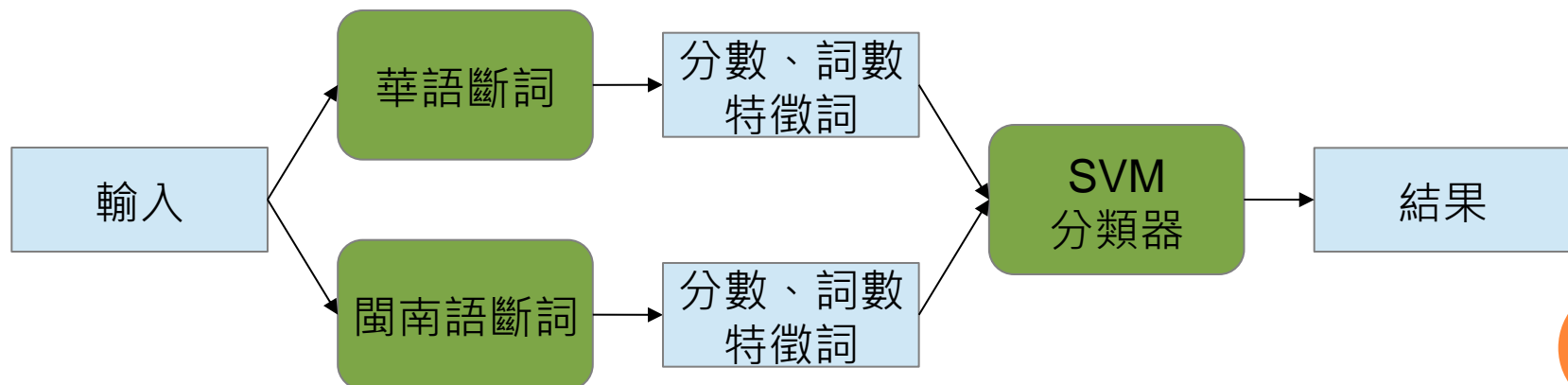
○ 定用詞數量 $n=7000$ ，特徵詞數量 $m=3000$

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

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



第四節：實驗結果

- 實驗一：斷詞效果
- 實驗二：校對的效果
- 實驗三：分類語言效果
- 實驗四：加入TGB的翻譯效果
- 實驗五：斷字恰斷詞的效果比較

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

實驗二 - 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
- 數位典藏329476句

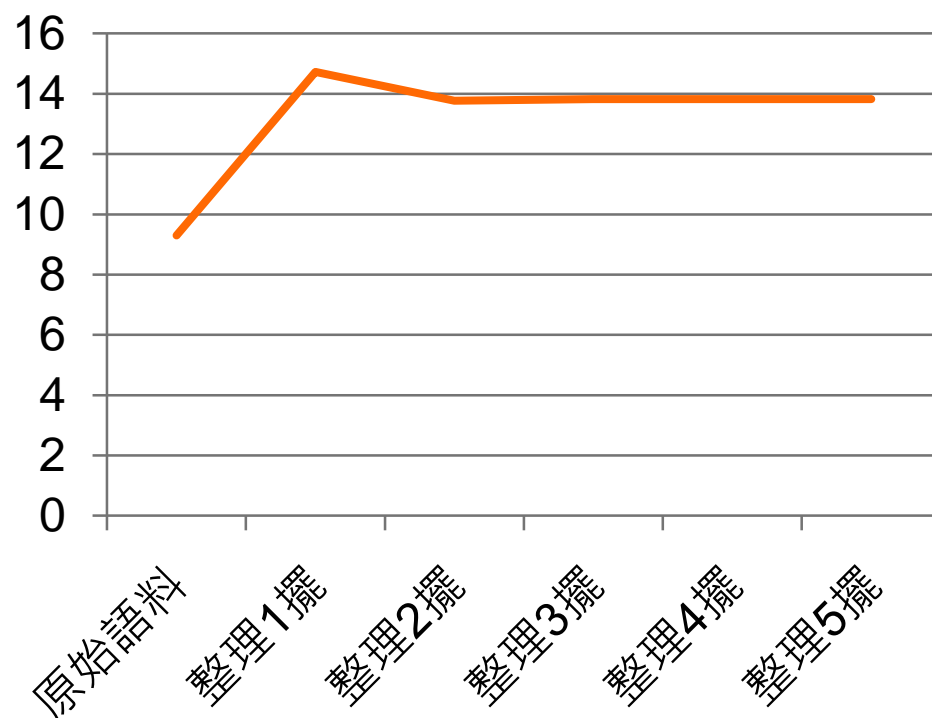
○ 試驗語料

- 教育部辭典例句8027句

○ 翻譯效果

- BLEU分數以詞為單位

BLEU



整理幾擺	0	1	2	3	4	5
BLEU分數	9.30	14.72	13.77	13.82	13.82	13.82

實驗三 - 分類語言

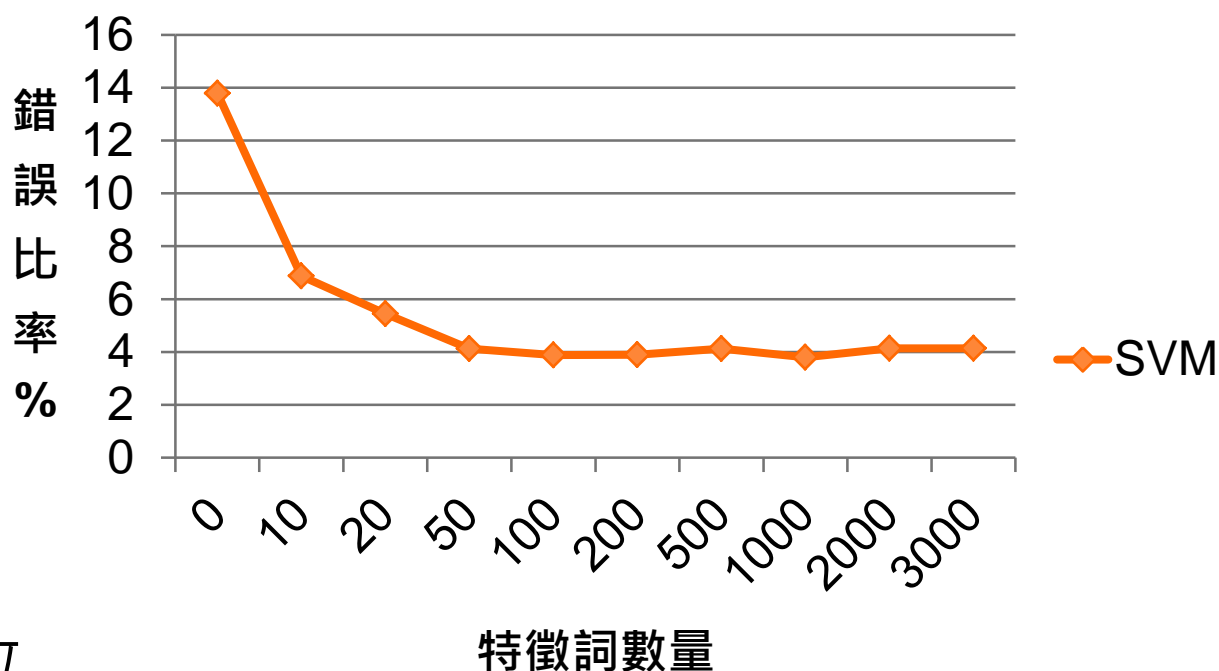
○ TGB通訊語料庫

○ 實驗參數

- 字詞數量
 - $k=4$
- 定用詞數量
 - $n=7000$
- 特徵詞數量
 - $m=0\sim3000$

○ 以段做辨識單位

分類結果



段數/詞數	總合	華語	閩南語
訓練	1000篇文章	8519/439436	9368/488844
試驗	179篇文章	2397/114901	1344/75282

實驗四 - 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
- 數位典藏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

實驗五 - 斷字恰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
- 數位典藏329476句
- TGB平行語料35025句

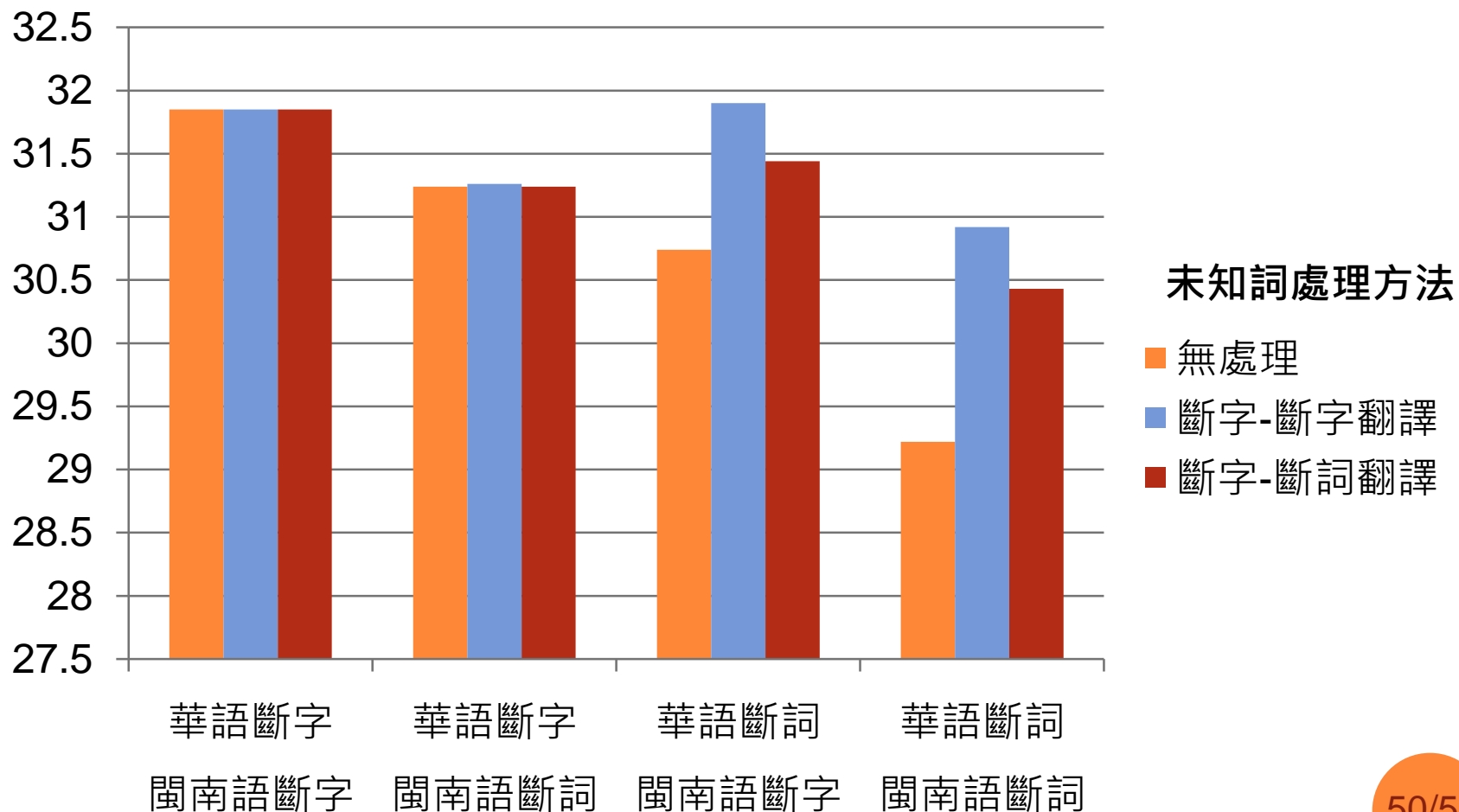
○ 試驗語料

- 教育部辭典平行例句8027句

○ 翻譯效果

- BLEU分數以字為單位

實驗五 - 斷字恰斷詞的效果比較實驗



第四節：結論與未來發展

○ 結論

- 拄好長度斷詞恰長詞優先差無濟
- 整理資訊無完整的語料庫，對翻譯有幫助
- 分類閩南語恰華語各50个特徵詞就夠用
- 華語有斷詞有較好
- 閩南語有斷詞顛倒無效果
 - 可能是斷詞的效果無好

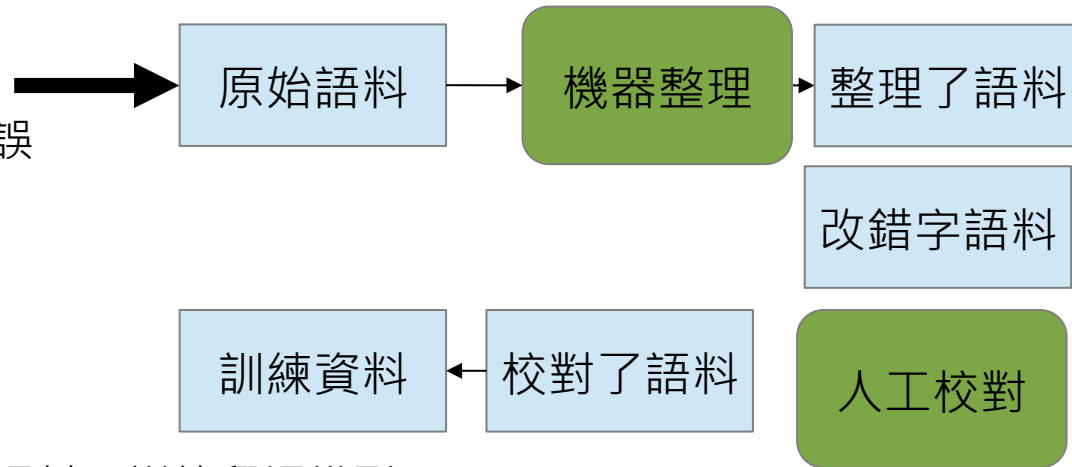
○ 未來發展

- 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

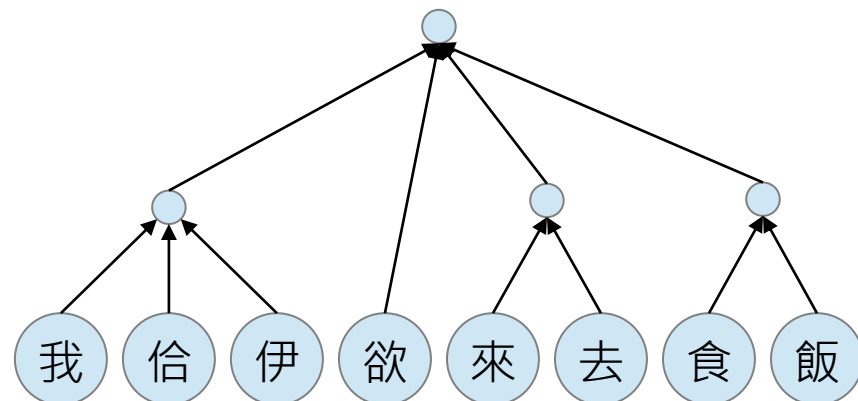
○ 自動校對語料

- 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」佢「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
 - 我佢伊欲來去食飯



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

第五節：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.

多謝逐家
