MRR

- MRR的計算方式就是對於每一品藏譯法華經品目，如果正確的漢譯對應品目，

  排序第一，得到1/1分；

  排序第二，得到1/2分；

  排序第三，得到1/3分；

  排序第四，得到1/4分；

  排序第五，得到1/5分；

  序位大於五，得到0分。

實驗設定

		專業藏漢佛學詞典		通用藏漢綜合詞典	
對列模式		使用停用詞表	不使用停用詞表	使用停用詞表	不使用停用詞表
Simple n-gram matching		NP+S	NP-S	NG+S	NG-S
Vector-space model	n-gram抽詞	VNP+S	VNP-S	VNG+S	VNG-S
	CKIP斷詞	VCP+S	VCP-S	VCG+S	VCG-S

N 表示n-gram
V 表示vector-space
C 代表使用中央研究院中文詞庫小組開發的CKIP斷詞系統
P 表示使用專業藏漢佛學詞典
G 表示使用通用藏漢綜合詞典
S 表示停用詞表

2012

結果與討論

DADHIC

實驗結果

實驗設定代號	MRR	Std. Dev.	Within Top 2	Out of Top 5	Worst Rank
NP+S	0.4583	0.3839	18	9	15
NP-S	0.3810	0.4041	11	11	27
NG+S	0.1994	0.2762	9	17	22
NG-S	0.1887	0.2699	8	17	24
VNP+S	0.4417	0.3813	15	8	27
VNP-S	0.4417	0.3813	15	8	27
VNG+S	0.1714	0.2752	3	16	25
VNG-S	0.1714	0.2752	3	16	25
VCP+S	0.6232	0.4400	19	7	22
VCP-S	0.6339	0.4538	19	8	22
VCG+S	0.2470	0.3825	6	17	28
VCG-S	0.2512	0.3854	7	17	28

一般討論 (1/2)

- 實驗設定VCP+S與VCP-S表現最好
  - 其MRR分別為0.6232與0.6339，平均而言，可以在前二個候選品目中找到對應品目。
- 實驗設定NP+S與NP-S表現不俗
  - 簡單的n-gram matching，採用專業佛學詞典，仍可在前三個候選品目找到對應品目，顯示適當的詞典資源，對於品目層次對應的重要性。

一般討論 (2/2)

- 藏譯《法華經》共有27品目，加上一藏譯跋，總計有28品目，漢譯《法華經》共有28品目，加上一漢譯跋，總計29品目。
- VCP+S與VCP-S各有19品可以在前二候選品目找到真正的漢譯品目，無法在前五候選品目找到真正的漢譯品分別為8品與7品
- 相對的，NP+S的實驗設定也能夠分別達到Within Top 2為18，Out of Top 5為9

專業詞典VS. 通用詞典

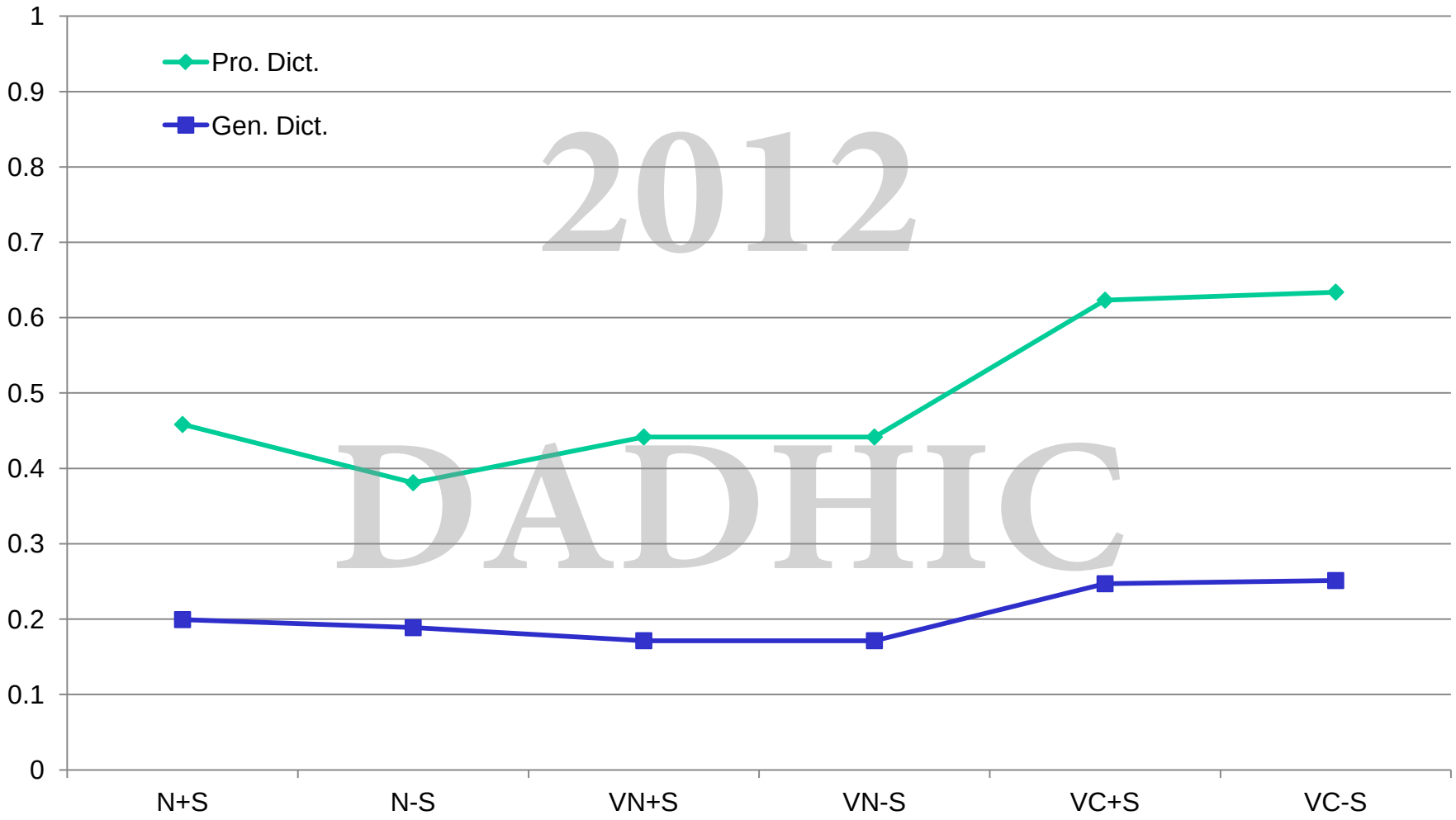
VC±S的實驗設定

- 專業佛學詞典可在前二個候選品目中，找到真正的對應品目
- 通用綜合詞典僅在前五個候選品目中，找到真正的對應品目

實驗結果依詞典不同分類

實驗設定	專業藏漢佛學詞典（Pro. Dict.）	通用藏漢綜合詞典（Gen. Dict.）
N+S	0.4583	0.1994
N-S	0.3810	0.1887
VN+S	0.4417	0.1714
VN-S	0.4417	0.1714
VC+S	0.6232	0.2470
VC-S	0.6339	0.2512
平均	0.4966	0.2049

專業詞典優於通用詞典的表現



Mann-Whitney U 檢定

$$U = n_1 * n_2 + \frac{n_1 * (n_1 + 1)}{2} - R_1$$

n_1為專業佛學詞典的數據個數

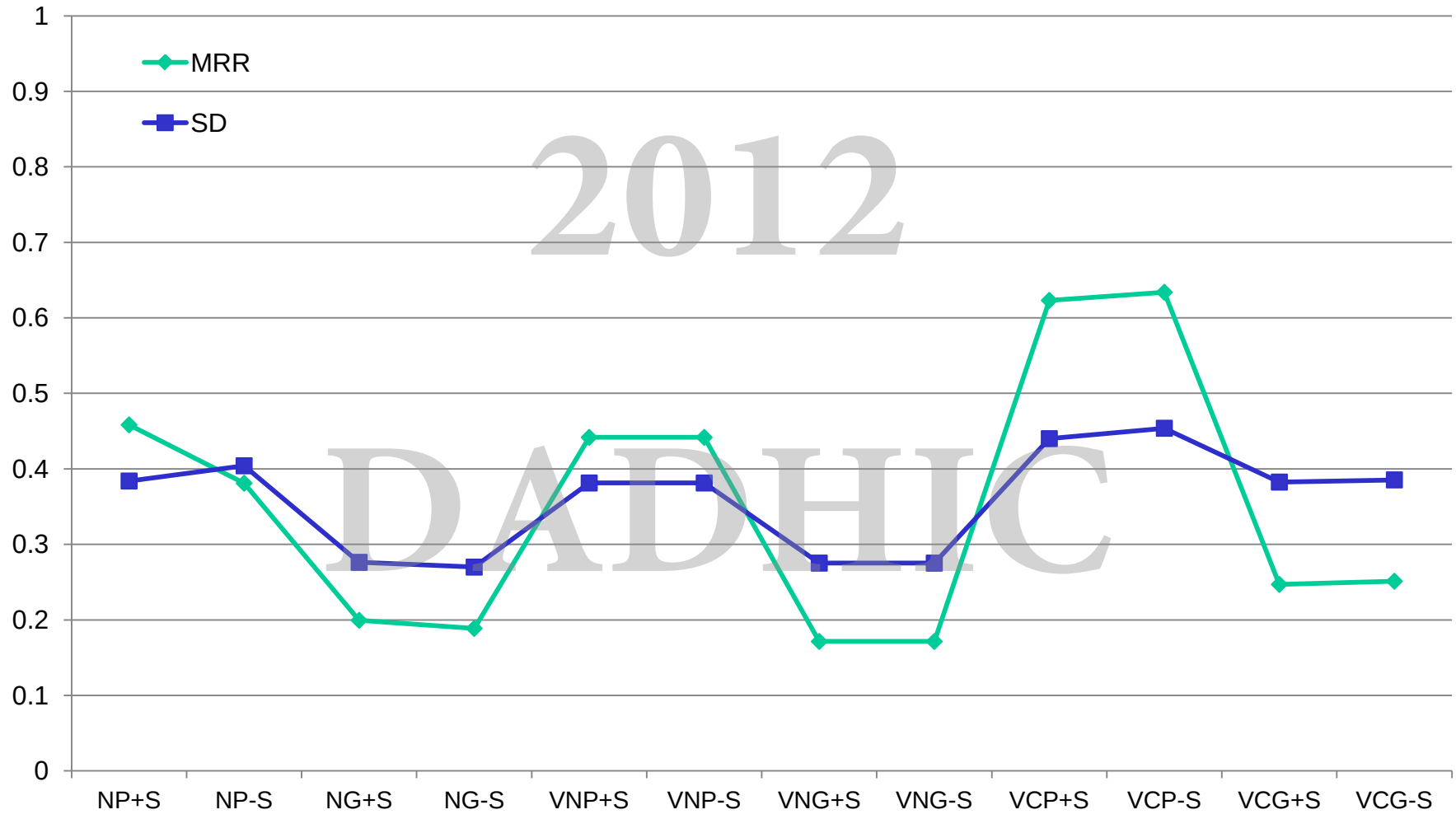
n_2為通用綜合詞典的數據個數

R_1為專業佛學詞典的數據等級和

檢定結果

- H_0:專業佛學詞典的表現與通用綜合詞典的表現相同
- H_1:專業佛學詞典的表現與通用綜合詞典的表現不同
- 依 Mann-Whitney U 檢定計算方式，得到R_1=57，U=0，p-value=0.0022 < α=0.05，因此拒絕虛無假設H_0，也就是專業佛學詞典與通用綜合詞典的表現差異達到統計上的顯著性。

各組MRR與對應標準差的遞變



結論 (1/2)

- 採用vector-space model，搭配CKIP中文斷詞處理、使用專業佛學詞典的實驗設定，可以在前二個候選品目找到真正的對應品目
- 簡單的n-gram matching方法，搭配專業佛學詞典，平均而言，也可以在前三個候選品目找到真正的對應品目

結論 (2/2)

- 實驗結果顯示專業藏漢佛學詞典對於處理不同譯本的對列，具有舉足輕重的角色
- 停用詞僅有在n-gram matching方法，有比較大的影響
- 不同語言譯本的佛教文獻品目層次的自動對列是可行的作法，可以有效降低文獻對勘研究初期工作需要的人力成本與時間成本

2012

敬請指教

DADHIC