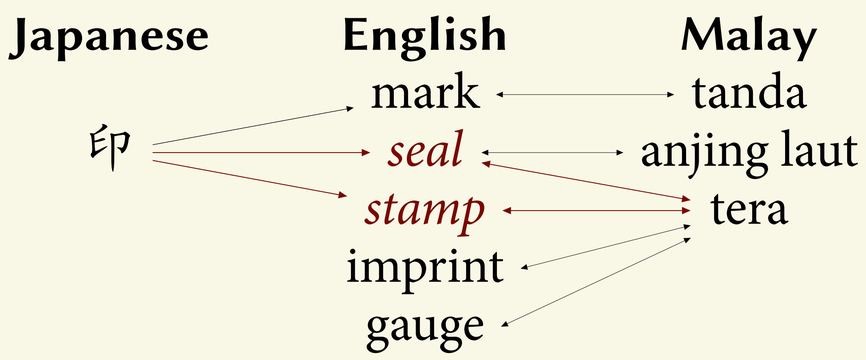
Low-Cost Construction of a Multilingual Lexicon from Bilingual Lists

Introduction

- Bilingual MRDs are good resources for building multilingual lexicons, but heterogeneous structures
- Lowest common denominator: list of source language item→ target language item(s)
- Multilingual lexicon construction using only simple bilingual lists (great for under-resourced language pairs)

One-time Inverse Consultation [1]

- Generates a bilingual lexicon for new language pair from existing bilingual lists
- JP–EN, EN–MS, MS–EN lexicons ⇒ JP–MS



$$\text{score}(\text{`tera'}) = 2 \times \frac{|\mathbb{E}_1 \cap \mathbb{E}_2|}{|\mathbb{E}_1| + |\mathbb{E}_2|} = 2 \times \frac{2}{3+4} = 0.57$$

∴ '印' ↔ 'tera' is most likely valid

Merging Translation Triples into Sets

- (Example: Malay–English–Chinese)
- Retain OTIC 'middle' language links
- For each 'head' language LI, discard triples with score $< \alpha X$ or score$^2 < \beta X$, where X = max score of all triples containing that LI

(garang, 凶猛) 0.143
(garang, ferocious, 凶猛)
(garang, fierce, 凶猛)

(garang, 激烈) 0.125
(garang, jazzy, 激烈)

(garang, 大胆) 0.111
~~(garang, bold, 大胆)~~

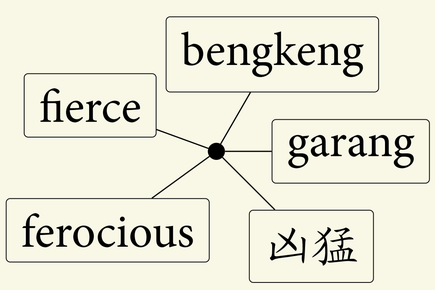
(garang, 黒体) 0.048
~~(garang, bold, 黒体)~~

(garang, 粗体) 0.048
~~(garang, bold, 粗体)~~

⋮

- Merge all triples with common bilingual pairs

(garang, ferocious, 凶猛)
(garang, fierce, 凶猛)
(bengkeng, fierce, 凶猛)

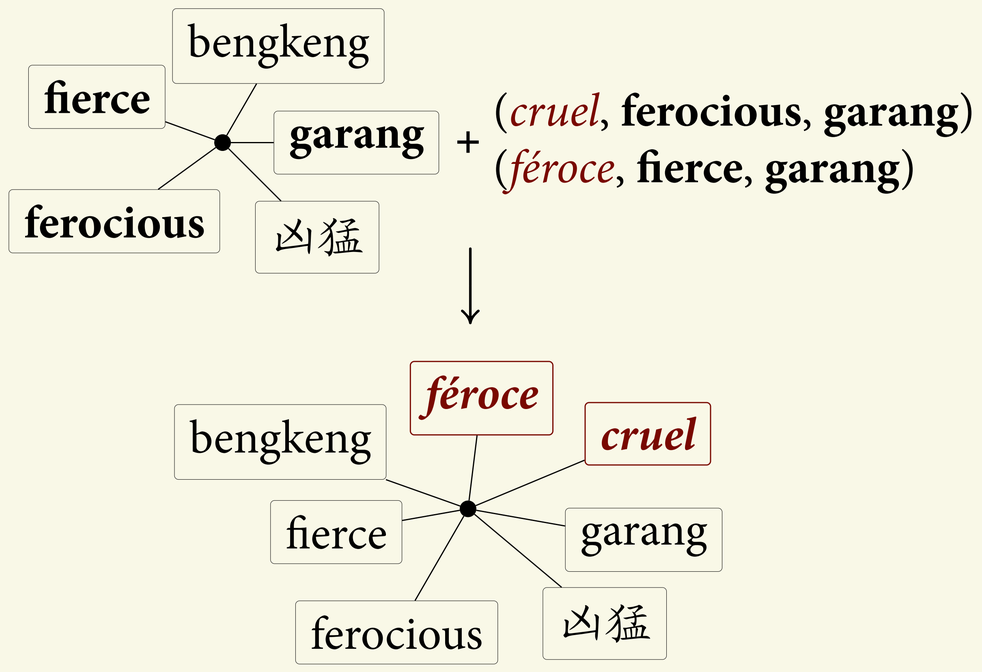


References

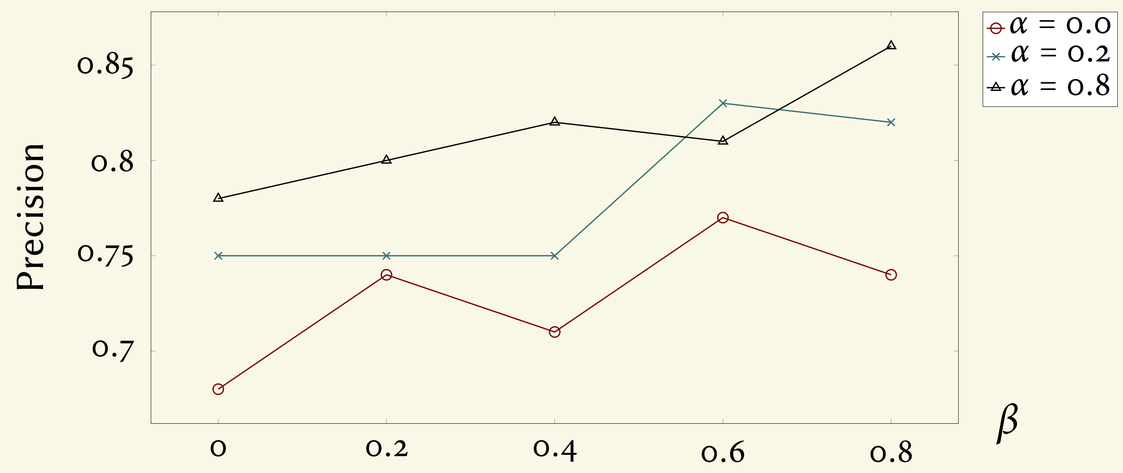
[1] F. Bond and K. Ogura. "Combining linguistic resources to create a machine-tractable Japanese–Malay dictionary". In: *Language Resources and Evaluation* 42 (2008), pp. 127–136.

Adding a New Language

- (Example: Malay–English–Chinese + French)
- Construct also French–English–Malay triples
- Add French members to existing M-E-C clusters with common English & Malay members



Precision of 100 Random Translation Sets



- Precision generally around 0.70–0.82; max 0.86

F_1 and Rand Index of Selected Translation Sets

- Evaluating accuracy of sets with polysemous 'middle' language members, e.g. 'plant', 'target'

Test word	Rand Index min	Rand Index max	F_1 min	F_1 max	Best accuracy when α	Best accuracy when β
'bank'	0.417	0.611	0.588	0.632	0.6	0.4
'plant'	0.818	0.927	0.809	0.913	0.6	0.2
'target'	0.821	1.000	0.902	1.000	0.4	0.2
'letter'	0.709	0.818	0.724	0.792	0.8	0.2

Discussion and Conclusion

- Low thresholds (α, β): more coverage; low precision
- High thresholds: good precision; low coverage
- $\alpha = 0.6, \beta = 0.2$ gives good trade-off between coverage, precision and recall
- Results are encouraging for such simple input data!
- Planned integration with an MT system

Lian Tze Lim liantze@gmail.com
Bali Ranaivo-Malançon ranaivo@mmu.edu.my
Enya Kong Tang enyakong@mmu.edu.my
NLP-SIG, Faculty of Information Technology, Multimedia University, Malaysia