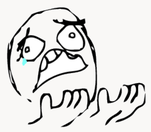
Context-Dependent Multilingual Lexical Lookup for Under-Resourced Languages

Input: Text Q = sequence of lexical items (LIs) $\{w_1, w_2, \ldots w_n\}$ **Output:** Ranked list of translations sets for each w_i

Motivation: Under-Resourced Languages

- WSD and translation selection typically require rich lexical knowledge and/or corpus resources
  - Gloss texts, subject codes, semantic networks
  - Corpora: (un)tagged; mono-/bilingual; aligned/comparable
- Whither under-resourced languages?
  - Lack of rich lexical resources
  - Small corpora

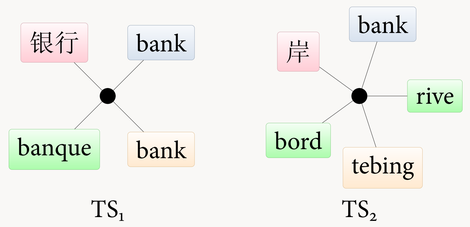


Comparable Bilingual Corpora

- **Idea:** leverage from *richer-resourced language pairs*
- Mine translation context data from 'rich' language pair
- Comparable corpora is easier to obtain
- E.g. concatenated English & Malay Wikipedia articles
- **62 993 documents, 67 499 terms**

Multilingual Lexcion

- **Translation sets**, like multilingual synsets
- Corresponding to coarse-grained concepts
- Member lexical items (LIs) from different languages



- **24 371 English, 13 226 Chinese, 35 640 Malay, 17 063 French, 14 687 Thai, 5629 Iban LIs**
- Uses translation context data from **rich language pair** for *entire translation set*

Extracting Translation Context Knowledge

- Document = bilingual article pair as bag-of-words
- Preprocess corpus (segment, stop word removal, lemmatise)
- Run latent semantic indexing (LSI) on corpus
  - 1000 factors, 45 minutes (MacBook Pro 2.3 GHz, 4 GB RAM)
  - One term vector for each LI in both languages
- $V(\text{TS}) = \sum(\text{all available term vectors of member LIs})$
- $\therefore V(\text{TS}_1) = V(\text{«bank»}_{\text{eng}}) + V(\text{«bank»}_{\text{msa}})$

$V(\text{TS}_2) = V(\text{«bank»}_{\text{eng}}) + V(\text{«tebing»}_{\text{msa}})$

Context-Dependent Lexical Lookup

- For input text Q, $V_Q = \sum V(w_i)$
- If Q language not in training corpus, $V_Q = \sum V(\text{all TS containing } w_i)$
- For each LI w_i in Q, $\text{TS}_{w_i} = \{t_1, \ldots, t_n\}$ set of all TS containing w_i
- TS_{w_i} sorted by $\text{CSim}(V_t, V_Q) = \dfrac{V_t \cdot V_Q}{|V_t| \times |V_Q|}$

Some Quick Results

- 80 text sentences: «bank»$_{\text{eng}}$, «plant»$_{\text{eng}}$, «kabinet»$_{\text{msa}}$, «mangga»$_{\text{msa}}$, «谷»$_{\text{zho}}$, «emperaja»$_{\text{iba}}$
- Strategies
  - `wiki-lsi` Proposed strategy
  - `base-freq` Baseline: most frequent translation
  - `goog-tr` Google Translate
- Metrics: Precision, Mean Reciprocal Rank

Strategy	Incl. Eng. & Iban		W/o Eng. & Iban	
	Precision	MRR	Precision	MRR
`wiki-lsi`	0.650	0.810	0.690	0.845
`base-freq`	0.550	0.771	0.524	0.762
`goog-tr`	0.797	0.812	0.690	0.708

Conclusions

- Trained on *bilingual* comparable corpus
- But can be used for *multilingual* inputs
- May not be highly accurate, but *fast, cheap* for under-resourced languages

Example with Iban Input (Top Ranked Translation Sets Shown)

Iban is a Bornean language with 600 000 speakers.

- Q = 'Lelaki nya tikah enggau emperaja iya, siko dayang ke ligung'

lelaki
zho: 男性
tha: ตัวผู้
fra: homme, mâle, masculin
msa: lelaki, jantan,
eng: male

tikah
zho: 结婚
tha: สมรส, ออกเรือน
fra: épouser, se marier
msa: menikahi, mengahwini
eng: marry, wed

emperaja
zho: 情人
tha: คู่ควง, ยอดรัก
msa: kekasih
eng: sweetheart

dayang
zho: 女孩子, 姑娘
tha: กัญญา, สาวน้อย
msa: pemudi, perawan, dara
eng: girl

ligung
zho: 可爱
tha: น่าเกลียดน่าชัง, น่ารักน่าชัง
fra: joli, mignon
msa: comel
eng: cute, pretty

Courtesy of LL-MAP
`llmap.org/languages/iba/static_map.html`

- Q = 'Udah ujan nya ngetu terbubuh, matahari enggau emperaja lalu ayan ba langit'

ujan
zho: 雨
tha: ฝนตก, พิรุณ, สายฝน
fra: flotte, pluie
msa: hujan
eng: rain

ngetu
zho: 停止
tha: ยั้ง, ขาดช่วง, แวะ
fra: arrêter, cesser, finir
msa: berhenti
eng: cease, quit, stop

matahari
zho: 太阳, 日
tha: ดวงตะวัน, พระอาทิตย์, ภากร
fra: solaire, soleil
msa: matahari
eng: solar, sun

emperaja
zho: 彩虹
tha: รุ้งกินน้ำ, สายรุ้ง, อินทรธนู
fra: arc-en-ciel
msa: pelangi
eng: rainbow

ayan
zho: 看得见
msa: terlihat
eng: perceptible, visible

langit
zho: 天空
tha: คคนางค์, ทิฆัมพร, เวหะ
fra: ciel
msa: langit
eng: sky, welkin

Lian Tze Lim * †
liantze@gmail.com
* SEST, KDU College Penang

Lay-Ki Soon †, **Tek Yong Lim** †
{lksoon,tylim}@mmu.edu.my
† FCI, Multimedia University

Enya Kong Tang ‡
enyakong1@gmail.com
‡ Linton University College

Bali Ranaivo-Malançon §
mbranaivo@fit.unimas.my
§ FCSIT, Universiti Malaysia Sarawak