▷ Requires large aligned thesaurus corpora (easier to acquire for specialised domains?)
▷ Cognate-based approach not applicable for language pairs that are not closely related

- Lafourcade [?]: compute contextual vectors for translation pairs based on gloss text and associated class labels from semantic hierarchy; compare vectors from different bilingual lexicons to detect synonymy
  - ▷ Resource requirements not available for all language pairs, costly task of assigning class labels

Contact

Lian Tze Lim	liantze@gmail.com
Bali Ranaivo-Malançon	ranaivo@mmu.edu.my
Enya Kong Tang	enyakong@mmu.edu.my

, Faculty of Information Technology
Multimedia University, Cyberjaya, Malaysia.
http://fit.mmu.edu.my/sig/nlp/

Low-Cost Construction of a Multilingual Lexicon from Bilingual Lists

Lian Tze Lim
Bali Ranaivo-Malançon
Enya Kong Tang

, Faculty of Information Technology
Multimedia University, Malaysia

Introduction

- Bilingual MRDs are good resources for building multilingual lexicons
- But MRDs have heterogeneous contents and structures
  - ▷ Not all contain rich information (gloss, domain) (Especially so for under-resourced languages)
  - ▷ Different structures (sense granularity, distinctions)
- Lowest common denominator: list of *source language item* $\rightarrow$ *target language item(s)*
- Construct multilingual lexicon using only bilingual lists

One-time Inverse Consultation [?]

- Generates a bilingual lexicon for a new language pair from existing bilingual lists
- Given bilingual lexicons L_1–L_2, L_2–L_3, L_3–L_2, generate bilingual lexicon L_1–L_3
- Example: JP–EN, EN–MS, MS–EN lexicons ⇒ JP–MS

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

$\therefore \leftrightarrow$ ‘tera’ is more likely to be valid

Merging Translation Triples into Sets

- Retain OTIC ‘middle’ language links

- For each 'head' language LI, filter only triples whose score exceed thresholds (See Algorithm 1)
- Merge all triples with common bilingual pairs
- Malay–English–Chinese example:
  - **MS–EN** Kamus Inggeris–Melayu untuk Penterjemah
  - **EN–ZH** XDict **ZH–EN** CC-CEDICT

Adding More Languages

- Construct L_1–L_2–L_4 triples
- Add L_4 members to existing L_1–L_2–L_3 clusters with common L_1 & L_2 members
- Example: Malay–English–Chinese + French, using 'ready-made' triples from FeM

Precision of 100 Random Translation Sets

- Precision increases with threshold parameters α and β
- Precision generally around 0.70–0.82; max 0.86
- Most false positives are not ranked at top of the list
- Many errors caused by incorrect POS assignments

F_1 and Rand Index of Selected Translation Sets

- False positives will frequently arise when 'middle' language members are polysemous, e.g. 'plant', 'target'
- Evaluate accuracy of selected sets with polysemous 'middle' language members

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{RI} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Algorithm 1: Generating trilingual translation chains

forall the *lexical items* $w_h \in L_1$ **do**
  $\mathbb{W}_m \leftarrow$ translations of w_h in L_2
  forall the $w_m \in \mathbb{W}_m$ **do**
    $\mathbb{W}_t \leftarrow$ translations of w_m in L_3
    forall the $w_t \in \mathbb{W}_t$ **do**
      Output a translation triple (w_h, w_m, w_t)
      $\mathbb{W}_{m_r} \leftarrow$ translations of w_t in L_2
      $\text{score}(w_h, w_m, w_t) \leftarrow$
      $$\sum_{w \in \mathbb{W}_m} \frac{|\text{common words in } w_{m_r} \in \mathbb{W}_{m_r} \text{ and } w|}{|\text{words in } w_{m_r} \in \mathbb{W}_{m_r}|}$$
    end
    $$\text{score}(w_h, w_t) \leftarrow 2 \times \frac{\sum_{w \in \mathbb{W}_m} \text{score}(w_h, w, w_t)}{|\mathbb{W}_m| + |\mathbb{W}_{m_r}|}$$
  end
  $X \leftarrow \max_{w_t \in \mathbb{W}_t} \text{score}(w_h, w_t)$
  forall the *distinct translation pairs* (w_h, w_t) **do**
    if *score*$(w_h, w_t) \geq \alpha X$ *or* $(\textit{score}(w_h, w_t))^2 \geq \beta X$ **then**
      Place $w_h \in L_1, w_m \in L_2, w_t \in L_3$ from all triples $(w_h, w_{\dots}, w_t)$ into same translation set
      Record $\text{score}(w_h, w_t)$ and $\text{score}(w_h, w_m, w_t)$
    else
      Discard all triples $(w_h, w_{\dots}, w_t)$
      // The sets are now grouped by (w_h, w_t)
    end
  end
end
Merge all sets containing triples with same (w_h, w_m)
Merge all sets containing triples with same (w_m, w_t)

Algorithm 2: Adding L_{k+1} to multilingual lexicon $\mathbb{L}$ of $\{L_1, L_2, \dots, L_k\}$

$T \leftarrow$ translation triples of L_{k+1}, L_m, L_n generated by Algorithm ?? where $L_m, L_n \in \{L_1, L_2, \dots, L_k\}$
forall the $(w_{L_m}, w_{L_n}, w_{L_{k+1}}) \in T$ **do**
  Add $w_{L_{k+1}}$ to all entries in $\mathbb{L}$ that contains both w_{L_m} and w_{L_n}
end

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

- F_1 and RI increases with α and β
- But may decrease when they are too high and reject valid members (false negatives)

Discussion

- Low thresholds (α, β): more coverage; low precision
- High thresholds: good precision; low coverage
- $\alpha \approx 0.6, \beta \approx 0.2$ gives good trade-off between coverage, precision and recall
- Results are encouraging for such simple input data! Especially suitable for under-resourced language pairs
- **Future plan:** Integrate multilingual lexicon into an MT system with WSD and user interaction features

Related Work

- Many multilingual lexicon projects [? ?]) aligned with Princeton WordNet [?]
  - Overly fine sense distinctions in Princeton WordNet
- Pan Lexicon [?]: compute context vectors of words from monolingual corpora of different languages, then grouping into translation sets by matching context vectors via bilingual lexicons
  - Sense distinctions derived from corpus evidence
  - Produces many translation sets that contain semantically related but not synonymous words, e.g. 'shoot' and 'bullet' (lower precision)
  - 44 % precision based on evaluators' opinions (75 % if inter-evaluator agreement is not required)
  - Does not handle multi-word expressions
- Markó, Schulz and Hahn [?] use cognate mappings to derive new translation pairs, validate by processing parallel corpora (medical domain)
  - Complex terms indexed on the level of sub-words e.g. 'pseudo⊕hypo⊕para⊕thyroid⊕ism'
  - 46 % accuracy for each language pair