CS657A: Information Retrieval
Cross-Lingual Information Retrieval

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs657/

2nd semester, 2021-22

Tue 1030-1145, Thu 1200-1315

Cross-Lingual IR

- Increasing usage of multiple languages
- Most are resource-poor
- Therefore, IR resources are hard to construct
- Cross-lingual IR helps
- Asking queries in one language and retrieving it from another

Typology of Models

- Type of *comparability*
- In a parallel corpus, words, sentences, and documents are *exact* translations of each other
- In a comparable corpus, words, sentences or documents are *similar* in some way
  - May be topics
- Type of *alignment*
- Word-level
- Sentence-level
- Document-level

	Parallel	Comparable
Word	Dictionary	Images
Sentence	Translation	Captions
Document	Epics/Religious books	Wikipedia/Query answers

- Most work has been done for parallel words and sentences

Context Counting Co-Occurrence Vectors

- In each language, each word encodes a vector of co-occurring words in context
- May use weights to denote frequency of co-occurrence
- A seed bi-lingual dictionary provides translated word pairs
  - माता, मा, মা
- Vectors of these word pairs are aligned to produce the most similarity
- This learns a vector space embedding for all words
- Vectors of other source words are subjected to the same transformation

Multi-Lingual Probabilistic Topic Modeling

- Two words in different languages with similar topic distributions are similar
- Mono-lingual LDA or LSA
- Assumes aligned document pairs
- May use a seed word pair dictionary to enforce similar distributions

Monolingual Mapping-based

- Train monolingual embeddings, and then align using parallel word pairs
- *Linear transformation matrix* $W^{s \rightarrow t}$
- Seed word pairs x^s, x^t
- Minimize mean squared error

$$\arg\min_{W} \Omega_{MSE} = \sum_{i=1}^{n} ||W.x_i^s - x_i^t||^2$$

- Canonical correlation analysis (CCA)
- Maximize correlation of projections of seed pairs

$$\arg\max_{u,v} \Omega_{CCA} = \sum_{i=1}^{n} \rho(x_i^s.u, x_i^t.v)$$

Pseudo-Multi-Lingual Merged Corpus

- Create a merged document containing both the languages
- Randomly replace words in s by translations in t
- Suppose word w^s has k_t translations $\{w^t\}$
- w^s is replaced by any of w^t with probability $1/(2k_t)$
- *Polysemy* can be handled by using context
- Choose the translation that is most similar to the context

$$w^t_{i*} = \arg\max_{\forall w^t_i} cos(w^s_i + w^s_{\forall c}, w^t_i)$$

Joint Space

- *Alignment matrix* $A^{s \to t}$ from s to t that defines which word in s can perform the same task as which word in t
- *Cross-lingual regularization term*

$$\Omega_{s \to t} = \sum_{i=1}^{V_s} {x_i^s}^T A^{s \to t} x_i^s$$

- Representation of translated word is closest to sum of context words

$$E(w_i^s, w_i^t) = -\Big(\sum_{j=-c}^{+c} {x_{i+j}^s}^T . T \Big) x_i^t$$

Word-Level Comparable Corpora

- Language grounding using images
- Similarity score of a pair of words is based on visual similarity of its associated image sets
- Better in augmenting text signals than working alone
- Parts-of-speech (POS) tags of words in the context
- POS distribution is likely to be same
- Assumption is that languages are close and written in the same manner
  - Bangla $\leftrightarrow$ Hindi, English $\leftrightarrow$ French

Sentence-Level Parallel Corpora

- Word-alignment matrices between word embeddings
- Choose alignment matrices that minimize

$$\Omega_{s \to t} + \Omega_{t \to s} = ||X^t - A^{s \to t} X^s||^2 + ||X^s - A^{t \to s} X^t||^2$$

- Compositional sentence models minimize error between sentence representations
- A sentence representation is sum of word representations

$$\min ||y^s - y^t||^2 \text{ where } y^s = \sum_{\forall i} x_i^s,\ y^t = \sum_{\forall j} x_j^t$$

- Bi-lingual auto-encoder models
- Re-construct target sentence from source sentence
- Multiple hidden layers
- At least one encoder layer and one decoder layer

Sentence-Level Comparable Corpora

- Captions of images
- Minimize image description pairs

$$\Omega^{image,s} + \Omega^{image,t}$$

- Can be used in addition to text

$$\Omega^{image,s} + \Omega^{image,t} + \Omega^{s,t}$$

Document-Level Corpora

- Document-level parallel data uses sentence-level parallel data only
- Sentence-alignment models can be extended by incorporating error term between the sentences, in addition to words

$$\alpha||y_k^s - y_k^t||^2 + (1-\alpha)\left(\frac{1}{m}\sum_{i=1}^{m} x_i^s + \frac{1}{n}\sum_{j=1}^{n} x_j^t\right)$$

- If sentences are not parallel, find transformation that maps paragraph vectors in d^s to those in d^t
- Concept or topic based models
- Two words are similar if their probabilistic topic distributions are similar

Merged Documents

- Learn *joint embedding* from merged documents

English	हिन्दी
Where the mind is without fear and the head is held high;	जहां चित्त भय से शून्य हो जहां हम गर्व से माथा ऊंचा करके चल सकें
...	...
Into that heaven of freedom, My Father, let my country awake.	उसी स्वातंत्र्य स्वर्ग में इस सोते हुए भारत को जगाओ

- *Merge and Shuffle*
- Concatenate documents from both languages
- Randomly shuffle words

  head उसी country गर्व चित्त ...स्वातंत्र्य high mind हम स्वर्ग heaven

- *Length-Ratio Shuffle*
- Alternate between two languages in ratio of document lengths

  Where the mind is जहां चित्त भय ...let my country awake भारत को जगाओ

- Can be extended to multiple languages