Cross-Lingual Semantic Similarity Measure for Comparable Articles

{ Motaz.Saad,David.Langlois,Kamel.Smaili }@loria.fr

Loria Laboratoire lorrain de recherche en informatique et ses applications

UNIVERSITÉ DE LORRAINE

1. Problem

Retrieve comparable documents

What are comparable documents?

- Cross-lingual documents
- Aligned at document level.
- Related to the same topic but they are not necessarily translations of each other.
- Have same content in terms of topic, but not in terms of syntax and sentences.
- Our work on Arabic-English comparable docs.

  $\Rightarrow$ **We use a cross-lingual similarity measure based on latent semantics**

3. Formulation

- Use same approach as Littman et al.[a].
- Monolingual term-document matrix X_a

$$X_a = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & \dots & d_n \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{pmatrix} \end{array} \quad (1)$$

- Cross-lingual term-document matrix X_{cl} (parallel or comparable corpus):

$$X_{cl} = \begin{array}{c} \\ t_1^a \\ t_2^a \\ \vdots \\ t_l^a \\ t_1^e \\ t_2^e \\ \vdots \\ t_m^e \end{array} \begin{array}{c} \begin{array}{cccc} d_1^u & d_2^u & \dots & d_n^u \end{array} \\ \begin{pmatrix} w_{11}^a & w_{12}^a & \dots & w_{1n}^a \\ w_{21}^a & w_{22}^a & \dots & w_{2n}^a \\ \vdots & \vdots & \ddots & \vdots \\ w_{l1}^a & w_{l2}^a & \dots & w_{ln}^a \\ w_{11}^e & w_{12}^e & \dots & w_{1n}^e \\ w_{21}^e & w_{22} & \dots & w_{2n}^e \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1}^e & w_{m2}^e & \dots & w_{mn}^e \end{pmatrix} \end{array} \quad (2)$$

- Each d_i^u in X_{cl} is the concatenation of the Arabic doc d_i^a and its corresponding English doc d_i^e.
- Weights w_{ij} in X_a & X_{cl} are the $tf\,idf$
- X_{cl} enables LSI to learn the relationship between terms which are semantically related in the same language or between two languages.
- Apply SVD on X_a & X_{cl}: USV^T

[a] Littman et al., "Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing", 1998

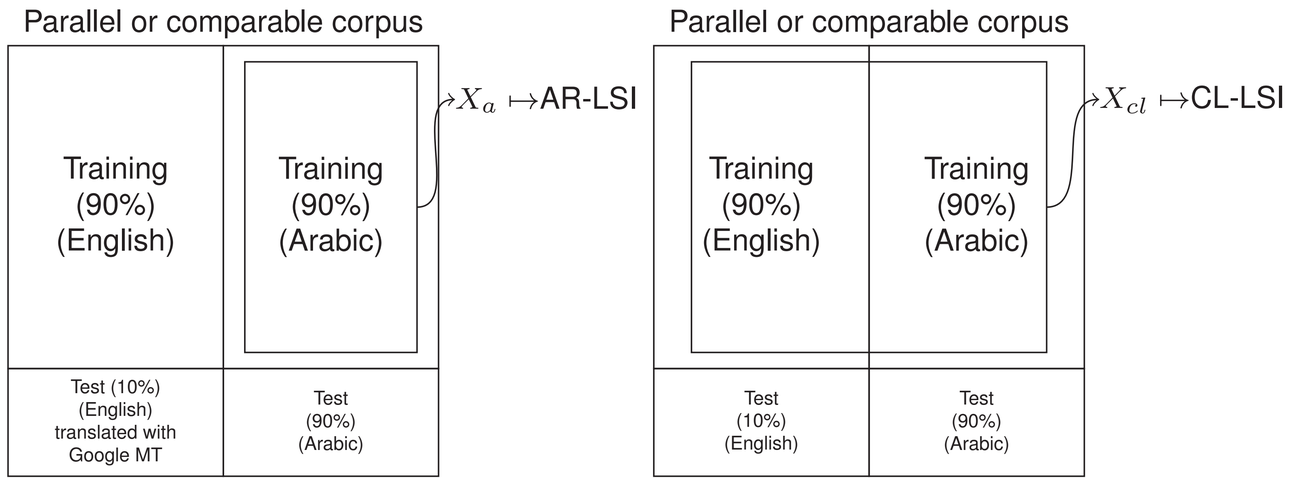
5. Results on parallel corpus

Retrieving parallel document mates using AR-LSI & CL-LSI

Corpus	Method	$R@1$	$R@5$	$R@10$
Newspapers				
AFP	AR-LSI	0.94	0.96	0.99
	CL-LSI	**0.97**	**0.99**	**0.99**
ANN	AR-LSI	0.80	0.91	0.94
	CL-LSI	**0.82**	**0.92**	**0.94**
ASB	AR-LSI	0.79	0.90	0.92
	CL-LSI	**0.85**	**0.92**	**0.97**
Medar	AR-LSI	0.56	0.76	0.81
	CL-LSI	**0.61**	**0.78**	**0.85**
NIST	AR-LSI	0.78	0.87	0.92
	CL-LSI	0.71	0.82	0.84
United Nations Resolutions				
UN	AR-LSI	0.97	1.00	1.00
	CL-LSI	**0.98**	0.99	**1.00**
Talks				
TED	AR-LSI	0.52	0.73	0.82
	CL-LSI	**0.60**	**0.83**	**0.92**
Movie Subtitles				
OST	AR-LSI	0.39	0.61	0.72
	CL-LSI	0.33	0.76	0.85

2. Cross-lingual similarity measure

- The cross-lingual similarity measure is based on Latent Semantic Indexing (LSI).
- Two ways:



4. Procedures

Algorithm 1: Retrieving Arabic documents using AR-LSI

Input: C_e: English corpus, C_a: Arabic corpus, n: number of docs to retrieve

$C'_e \leftarrow \varnothing$; $C'_a \leftarrow \varnothing$;
foreach *doc* a_j *in* C_a **do**
  $a'_j \leftarrow a_j^t US^{-1}$; put a'_j in C'_a;
foreach *doc* e_i *in* C_e **do**
  $a_{e_i} \leftarrow$ `translate`(e_i); $a'_{e_i} \leftarrow a_{e_i}^t US^{-1}$;
  $R \leftarrow$ `retrieve`(a'_{e_i}, C'_a, n) `// retrieve top-n similar docs to` e'_i `from` C'_a
  `evaluate`(R) `// check if` a'_i `in` R

Algorithm 2: Retrieving Arabic documents using CL-LSI

Input: C_e: English corpus, C_a: Arabic corpus, n: number of docs to retrieve

$C'_e \leftarrow \varnothing$; $C'_a \leftarrow \varnothing$;
foreach *doc* a_j *in* C_a **do**
  $a'_j \leftarrow a_j^t US^{-1}$; put a'_j in C'_a;
foreach *doc* e_i *in* C_e **do**
  $e'_i \leftarrow e_i^t US^{-1}$;
  $R \leftarrow$ `retrieve`(e'_i, C'_a, n) `// retrieve top-n similar docs to` e'_i `from` C'_a
  `evaluate`(R) `// check if` a'_i `in` R

Procedure retrieve(d_{s_i}, C_t, n)

Input: d_{s_i}: source doc, C_t: target corpus, n: number of docs to retrieve

$R \leftarrow \varnothing$; `// a list of retrieved docs`
foreach *doc* d_{t_j} *in* C_t **do**
  $sim \leftarrow \cos(d_{s_i}, d_{t_j})$; put (j, sim) in R;
  Sort R in descending order according to sim values;
return *top* n *elements of* R;

6. Results on comparable corpus

Retrieving comparable documents using CL-LSI

Corpus	$R@1$	$R@5$	$R@10$
Wikipedia	0.42	0.84	0.94
Euro-news	0.84	0.99	1.0

Comparing corpora using CL-LSI

Corpus	parallel	comparable
$\mathrm{avg}(\cos)$	0.53	0.46

7. Conclusion

- CL-LSI is competitive to monolingual LSI
- The advantage of CL-LSI is that it does not need machine translation !
- CL-LSI is language independent approach
- Future work: use CL-LSI to align English-Arabic comparable documents.