

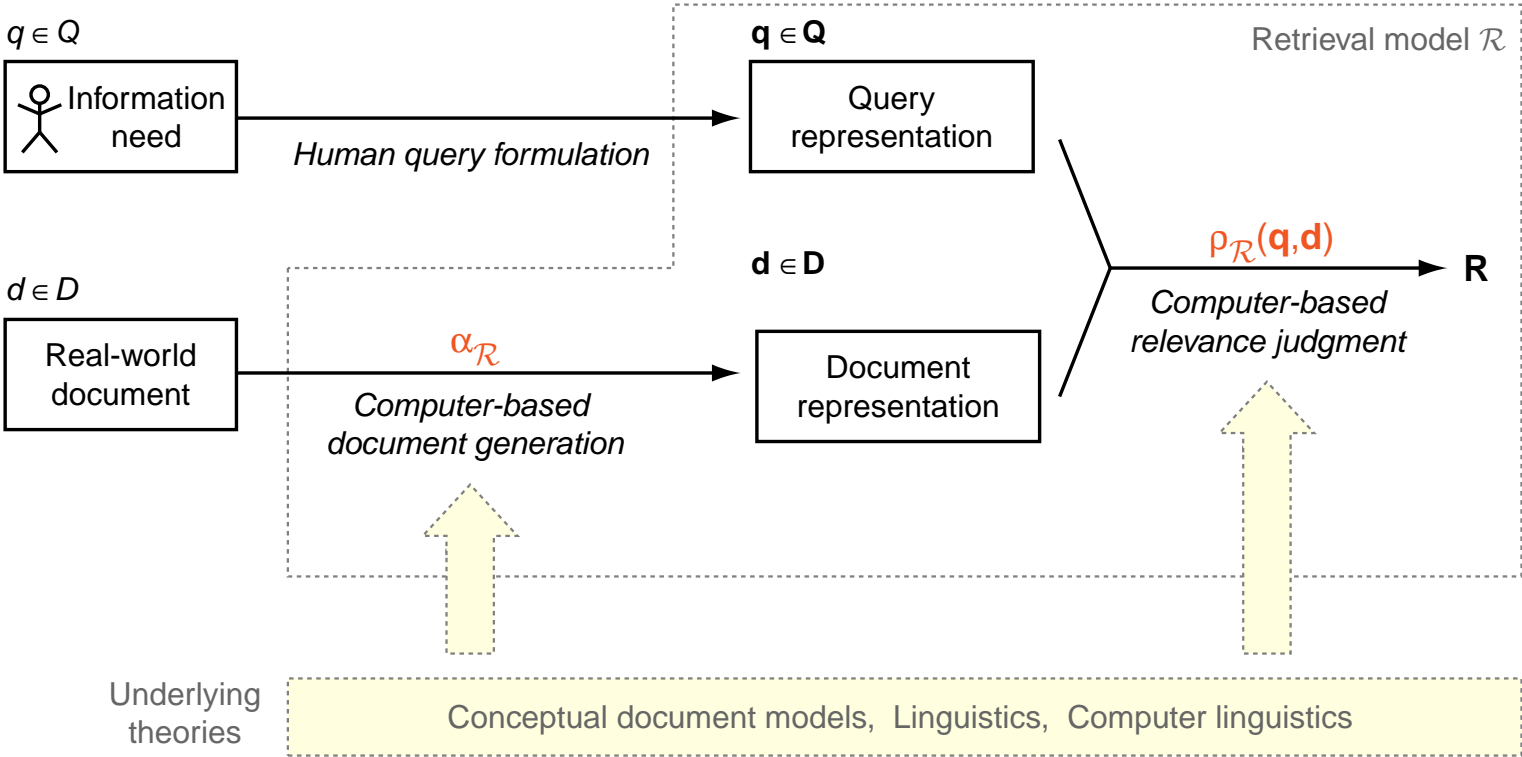
Cross-Language Explicit Semantic Analysis

Nedim Lipka Maik Anderka Benno Stein
Bauhaus University Weimar
www.webis.de

Outline

- ❑ Retrieval Models
- ❑ The CL-ESA Retrieval Model
- ❑ CL-ESA at TEL@CLEF 2009
- ❑ Formalization of CL-ESA

Retrieval Models

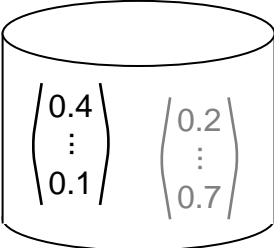


The CL-ESA Retrieval Model

Explicit Semantic Analysis, ESA [Gabrilovich/Markovitch 2007]

The CL-ESA Retrieval Model

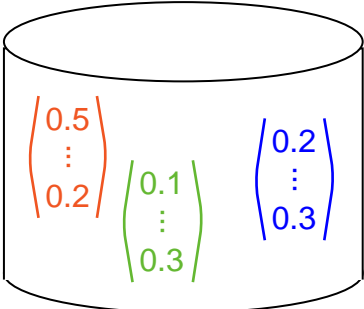
Explicit Semantic Analysis



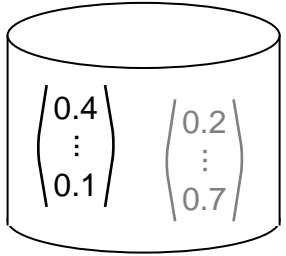
Document collection D

The CL-ESA Retrieval Model

Explicit Semantic Analysis



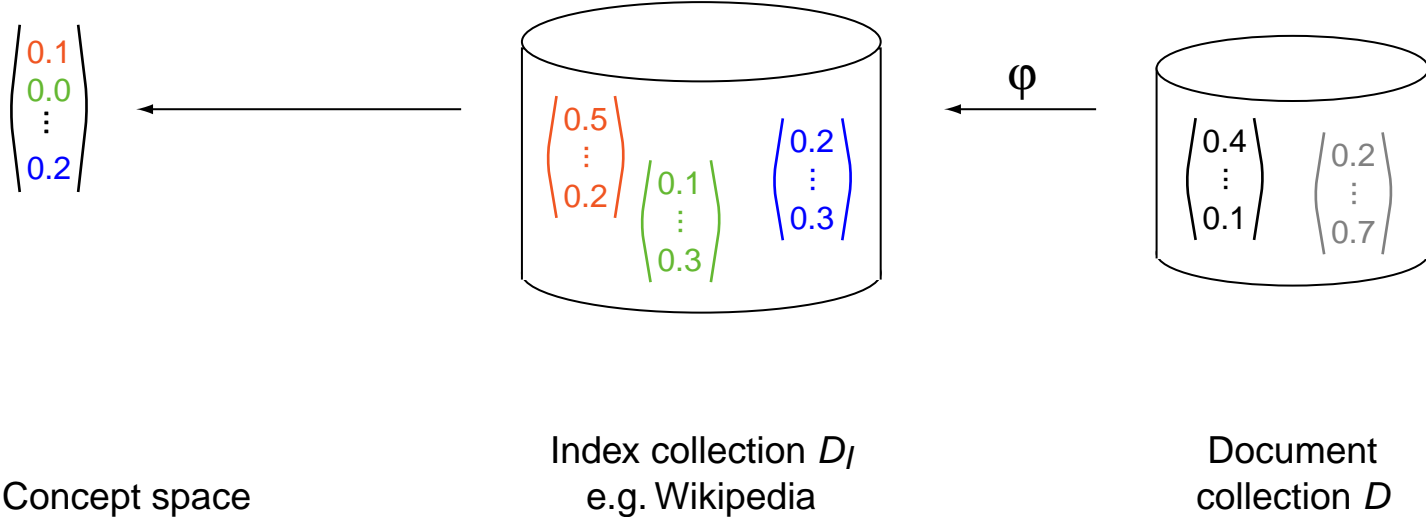
Index collection D_I
e.g. Wikipedia



Document collection D

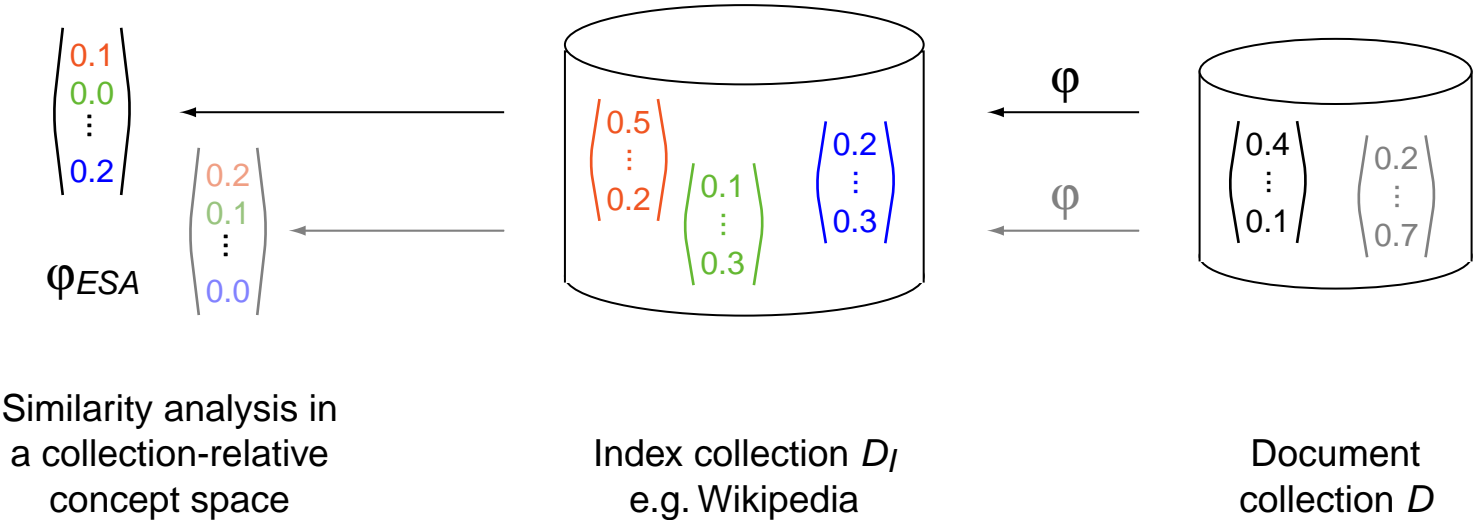
The CL-ESA Retrieval Model

Explicit Semantic Analysis



The CL-ESA Retrieval Model

Explicit Semantic Analysis



Ranking:

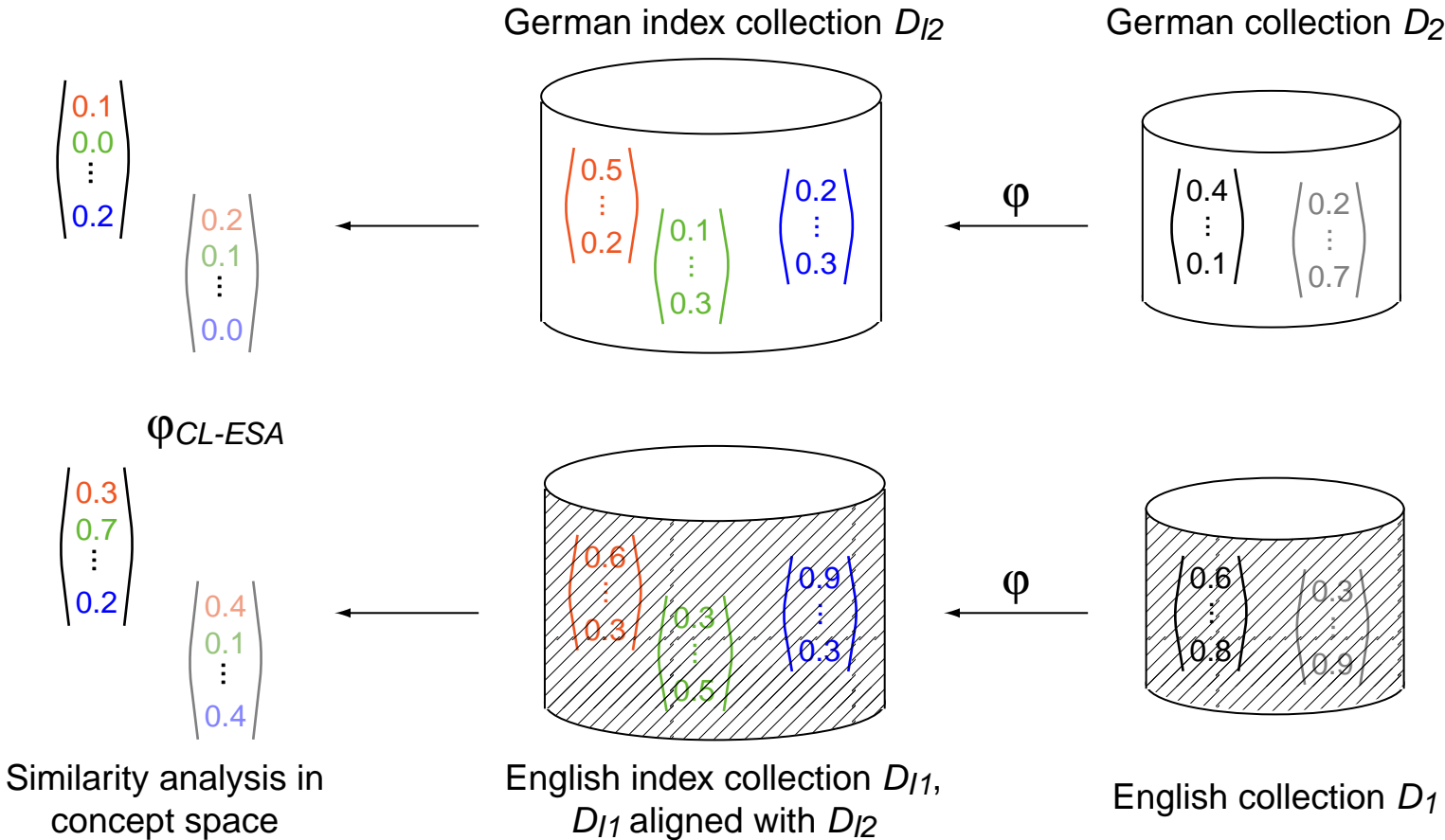
$$d^* = \operatorname{argmax}_{d \in D} \varphi_{ESA}(q, d),$$

where

$$\varphi_{ESA}(q, d) := \varphi(\mathbf{q}_{|D_I}, \mathbf{d}_{|D_I})$$

The CL-ESA Retrieval Model

Cross-Language Explicit Semantic Analysis



CL-ESA at TEL@CLEF 2009

Setting

Index collection:

- ❑ Wikipedia snapshot March 2009
- ❑ 169000 articles per language
- ❑ 3 index collections
- ❑ Query representation: title + description
- ❑ Document representation: title + subject + alternative

CL-ESA at TEL@CLEF 2009

Setting

Index collection:

- ❑ Wikipedia snapshot March 2009
- ❑ 169000 articles per language
- ❑ 3 index collections
- ❑ Query representation: title + description
- ❑ Document representation: title + subject + alternative

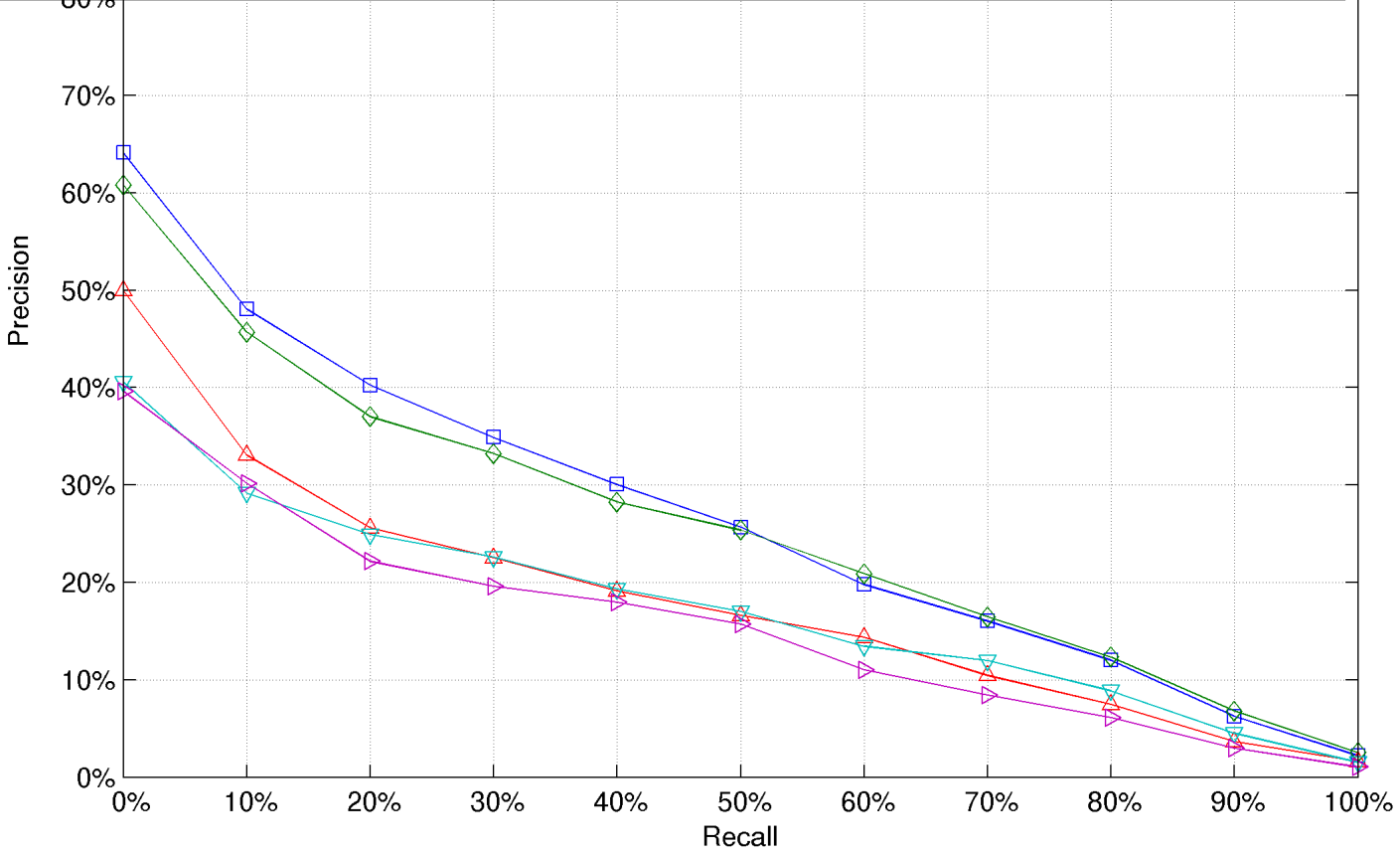
Difficulties at TEL@CLEF:

- ❑ Selecting the correct index collection. (language detection needed)
- ❑ Correct index collection not always available.
- ❑ Fields title, subject, and alternative not always share the same language.

CL-ESA at TEL@CLEF 2009

Ad-Hoc TEL Bilingual French Task Top 5 Participants - Standard Recall Levels vs Mean Interpolated Precision

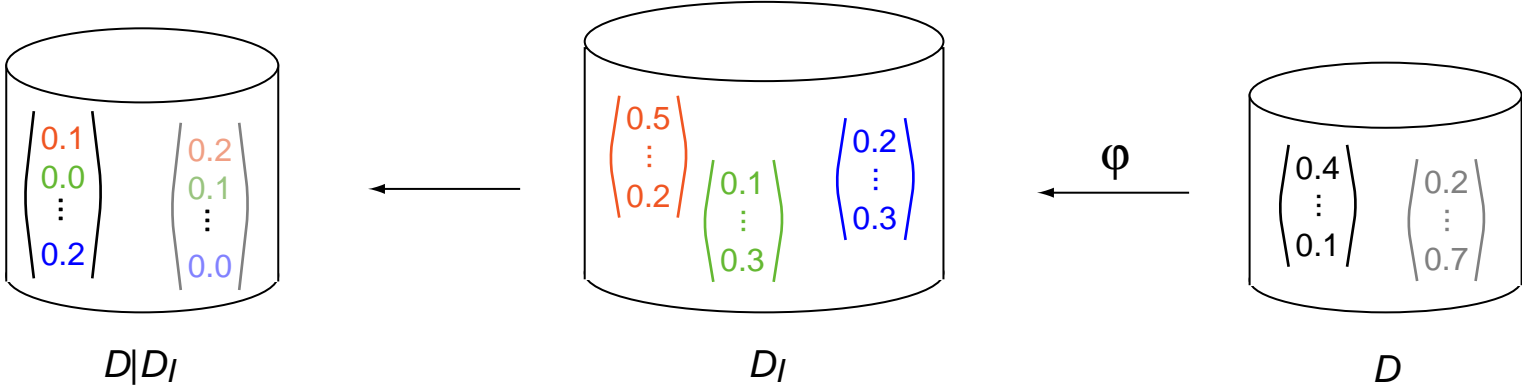
- chemnitz [Experiment CUT_24_BILI_EN2FR_MERGED_LANG_SPEC_REF_CUT_17; MAP 25.57%; Not Pooled]
- karlsruhe [Experiment EN_INDEXBL; MAP 24.62%; Not Pooled]
- cheshire [Experiment BIENFRT2FB; MAP 16.77%; Not Pooled]
- trinity [Experiment TCDDEFRRUN2; MAP 16.33%; Not Pooled]
- weimar [Experiment CLESA169283ENINFR; MAP 14.51%; Pooled]



Formalization of CL-ESA

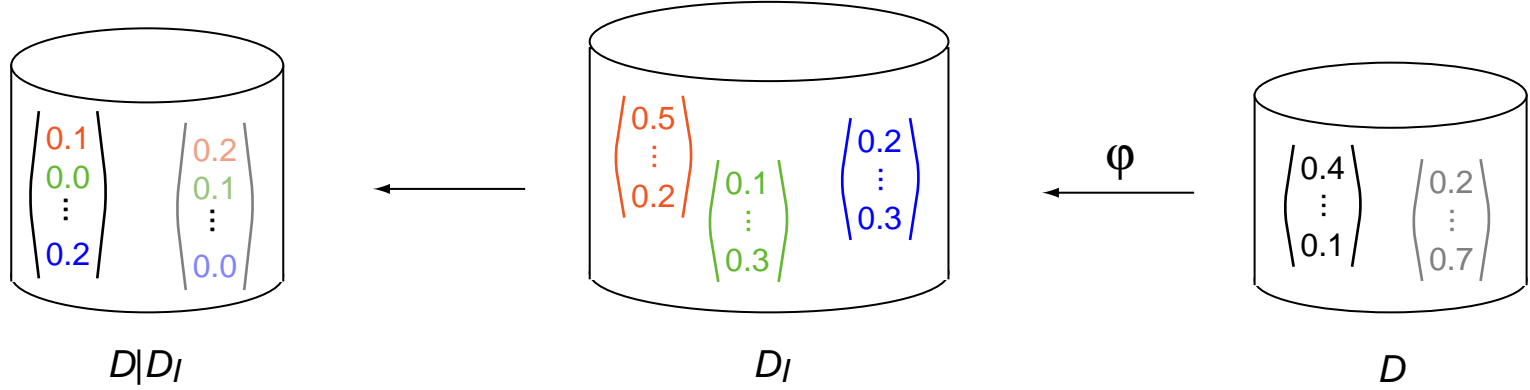
Formalization of CL-ESA

ESA

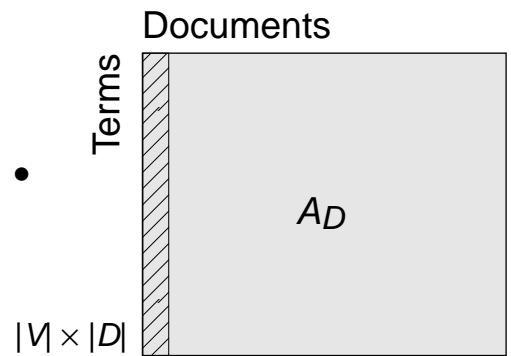
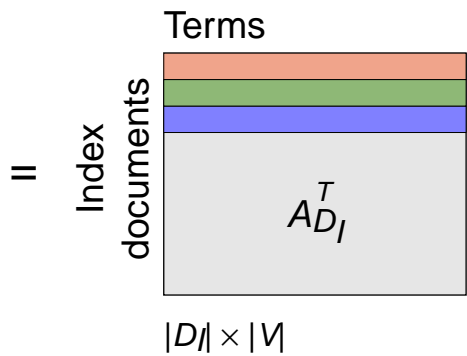
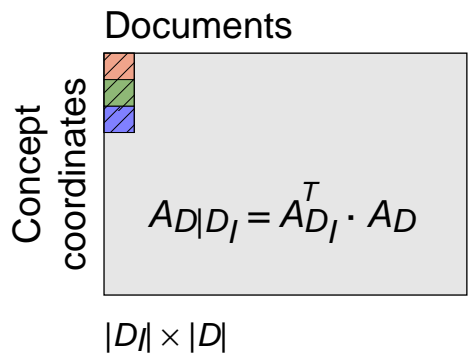


Formalization of CL-ESA

ESA



$$A_{D|D_I} = A_{D_I}^T \cdot A_D$$



Formalization of CL-ESA

CL-ESA

$$\begin{aligned}\varphi_{CL-ESA}(q, d) &= \varphi(\mathbf{q}|_{D_{I_1}}, \mathbf{d}|_{D_{I_2}}), \quad \text{with } D_{I_1}, D_{I_2} \text{ aligned} \\ &= \varphi(A_{D_{I_1}}^T \cdot \mathbf{q}, A_{D_{I_2}}^T \cdot \mathbf{d}) \\ &= nf (A_{D_{I_1}}^T \cdot \mathbf{q})^T \cdot A_{D_{I_2}}^T \cdot \mathbf{d} \\ &= nf \mathbf{q}^T \cdot A_{D_{I_1}} \cdot A_{D_{I_2}}^T \cdot \mathbf{d}\end{aligned}$$

Formalization of CL-ESA

CL-ESA

$$\begin{aligned}\varphi_{CL-ESA}(q, d) &= \varphi(\mathbf{q}_{|D_{I_1}}, \mathbf{d}_{|D_{I_2}}), \quad \text{with } D_{I_1}, D_{I_2} \text{ aligned} \\ &= \varphi(A_{D_{I_1}}^T \cdot \mathbf{q}, A_{D_{I_2}}^T \cdot \mathbf{d}) \\ &= nf (A_{D_{I_1}}^T \cdot \mathbf{q})^T \cdot A_{D_{I_2}}^T \cdot \mathbf{d} \\ &= nf \mathbf{q}^T \cdot A_{D_{I_1}} \cdot A_{D_{I_2}}^T \cdot \mathbf{d} \quad \sim \text{Cross language term co-occurrence} \\ &= nf \mathbf{q}^T \cdot G_{L_1, L_2} \cdot \mathbf{d}\end{aligned}$$

Formalization of CL-ESA

CL-ESA

$$\begin{aligned}\varphi_{CL-ESA}(q, d) &= \varphi(\mathbf{q}_{|D_{I_1}}, \mathbf{d}_{|D_{I_2}}), \quad \text{with } D_{I_1}, D_{I_2} \text{ aligned} \\ &= \varphi(A_{D_{I_1}}^T \cdot \mathbf{q}, A_{D_{I_2}}^T \cdot \mathbf{d}) \\ &= nf (A_{D_{I_1}}^T \cdot \mathbf{q})^T \cdot A_{D_{I_2}}^T \cdot \mathbf{d} \\ &= nf \mathbf{q}^T \cdot A_{D_{I_1}} \cdot A_{D_{I_2}}^T \cdot \mathbf{d} \quad \sim \text{Cross language term co-occurrence} \\ &= nf \underbrace{\mathbf{q}^T \cdot G_{L_1, L_2}}_{\substack{\text{Query} \\ \text{translation}}} \cdot \mathbf{d}\end{aligned}$$

Outlook

1. Consideration of more index collections
2. Better language detection
3. Detailed analysis of document fields

