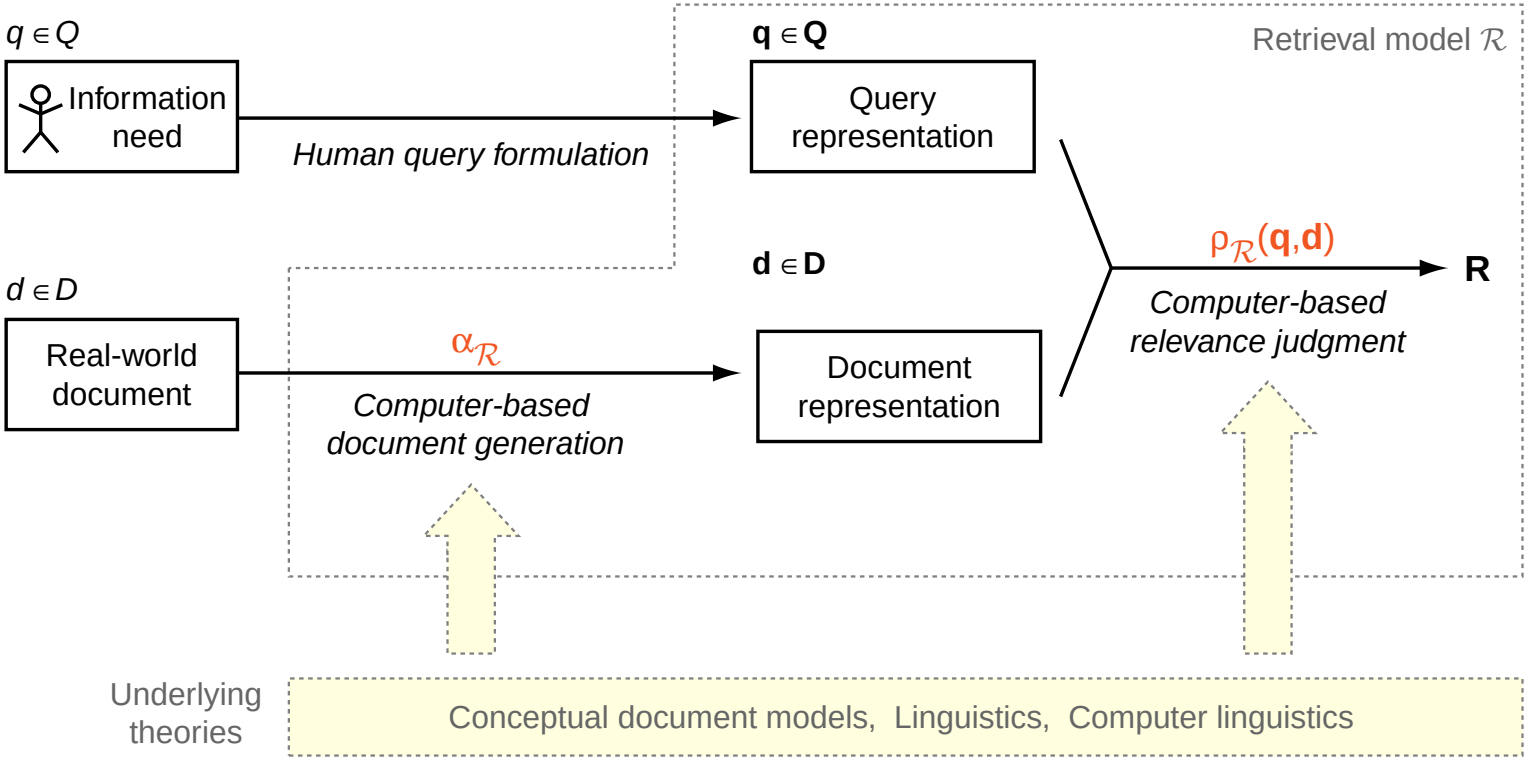
Cross-Language Explicit Semantic Analysis

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

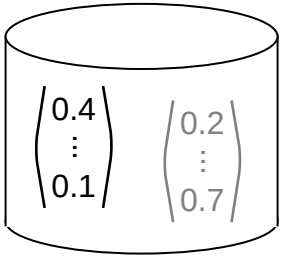


The CL-ESA Retrieval Model

Explicit Semantic Analysis, ESA [Gabrilovich/Markovitch 2007]

The CL-ESA Retrieval Model

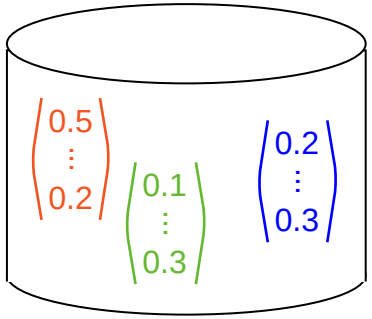
Explicit Semantic Analysis



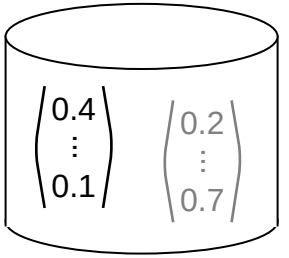
Document collection *D*

The CL-ESA Retrieval Model

Explicit Semantic Analysis



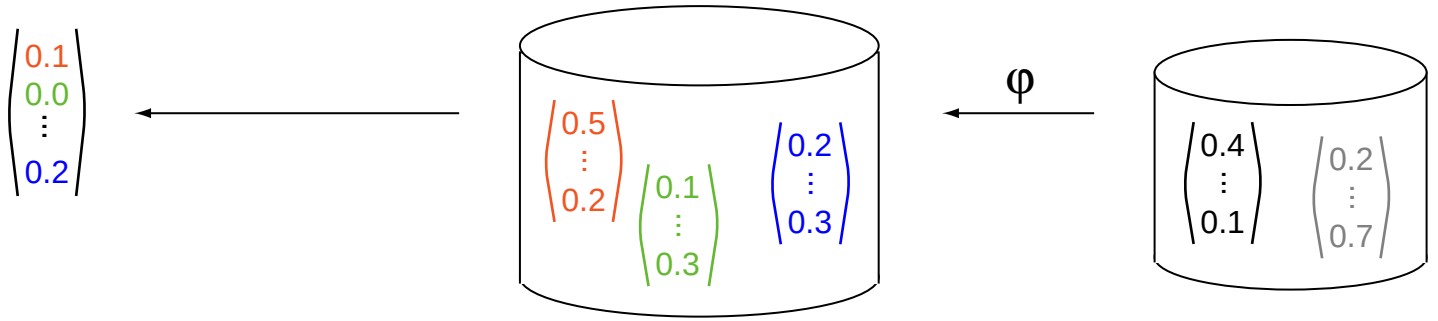
Index collection D_I
e.g. Wikipedia



Document
collection D

The CL-ESA Retrieval Model

Explicit Semantic Analysis



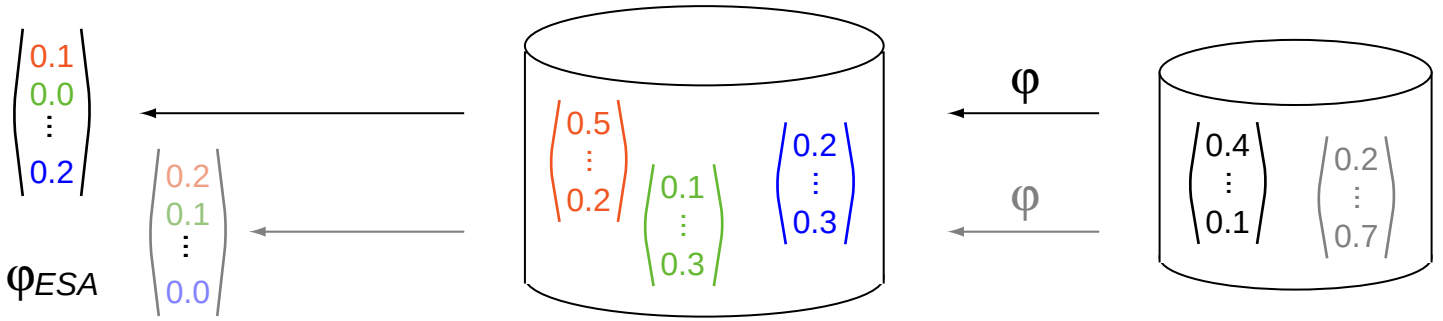
Concept space

Index collection D_I
e.g. Wikipedia

Document
collection D

The CL-ESA Retrieval Model

Explicit Semantic Analysis



Similarity analysis in a collection-relative concept space

Index collection D_I e.g. Wikipedia

Document collection D

Ranking:

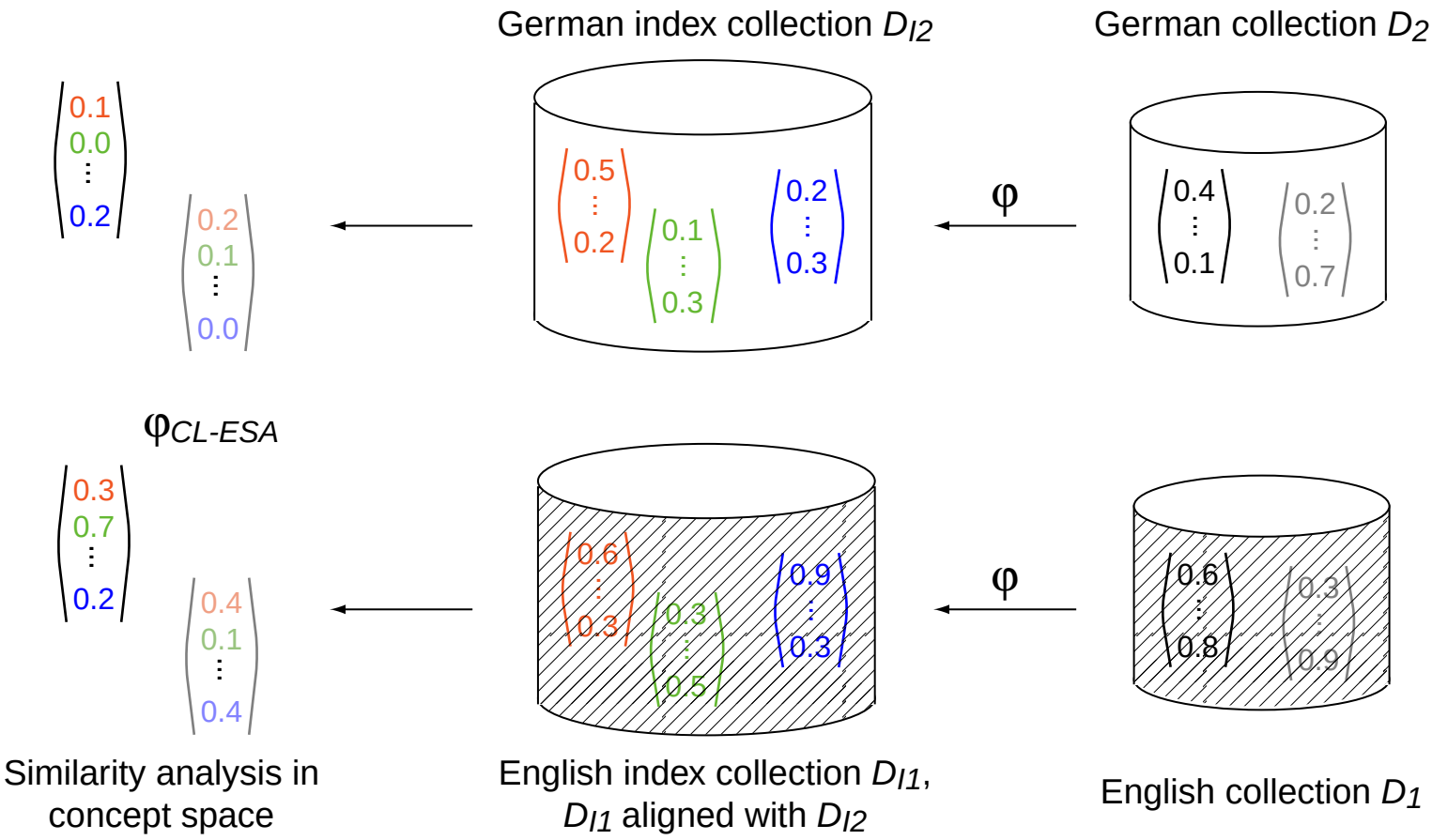
$$d^* = \text{argmax}_{d \in D} \ \varphi_{ESA}(q, d),$$

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

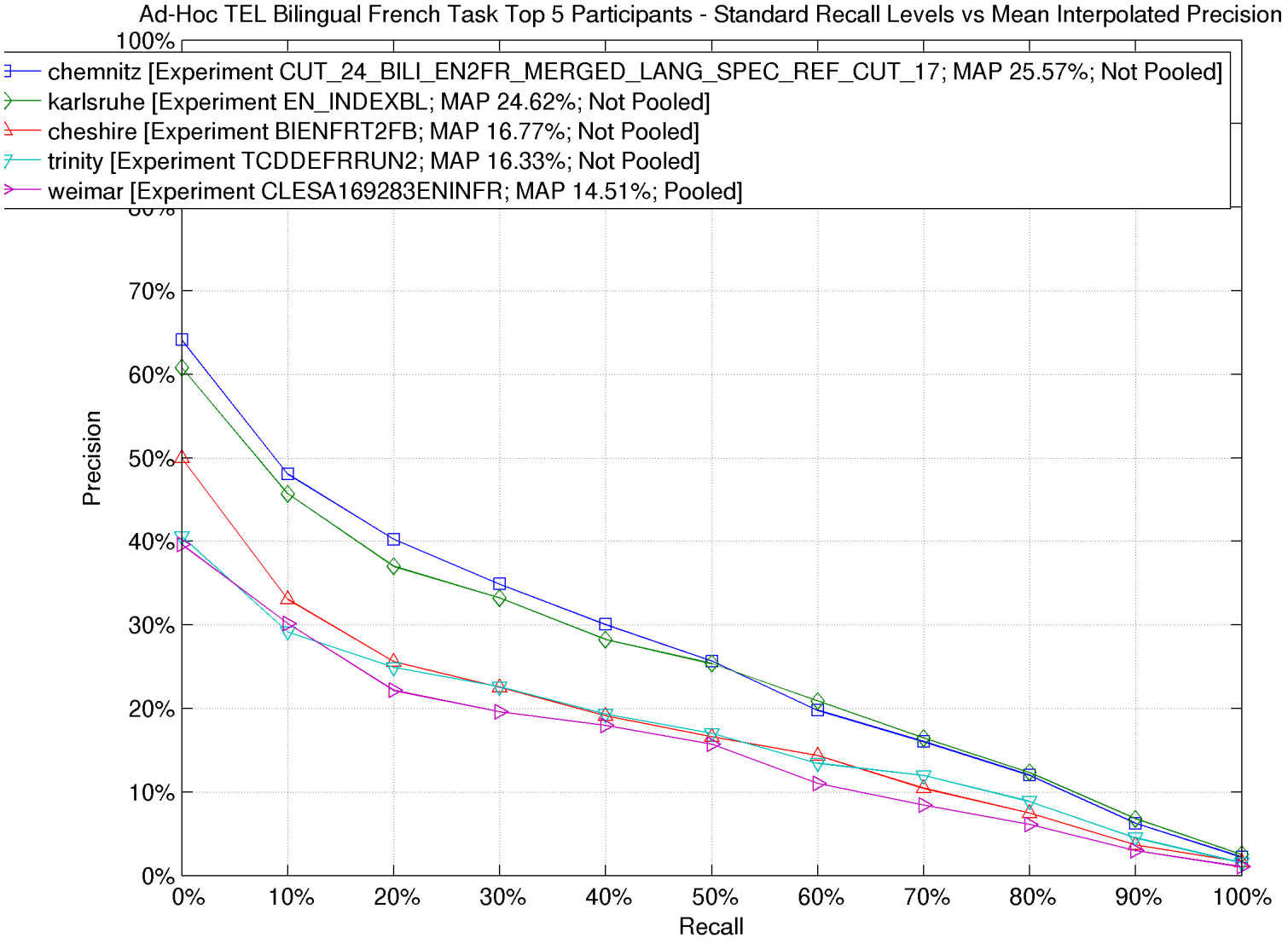
Index collection:

- 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.

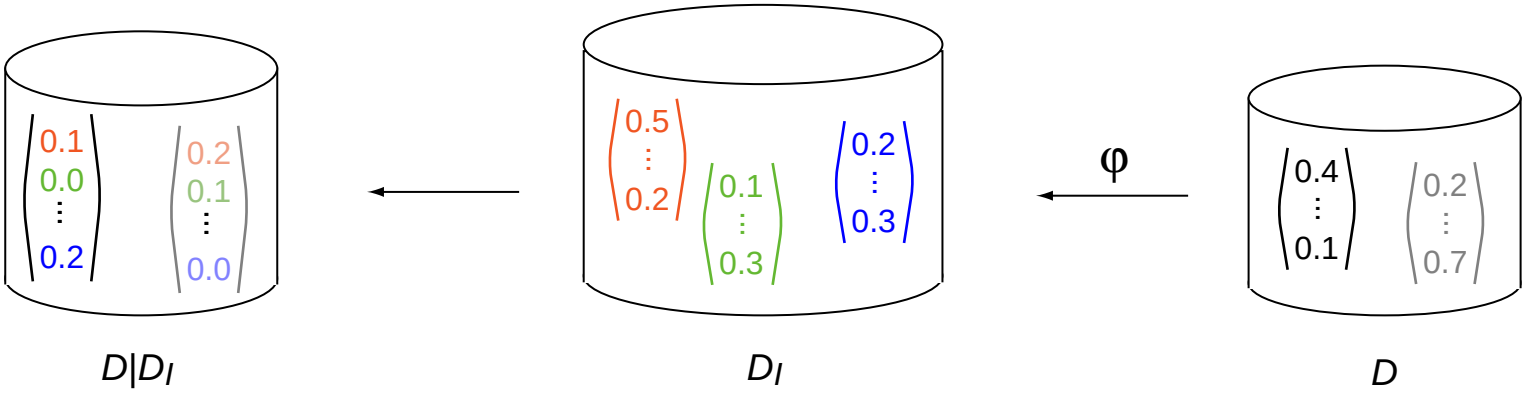
CL-ESA at TEL@CLEF 2009



Formalization of CL-ESA

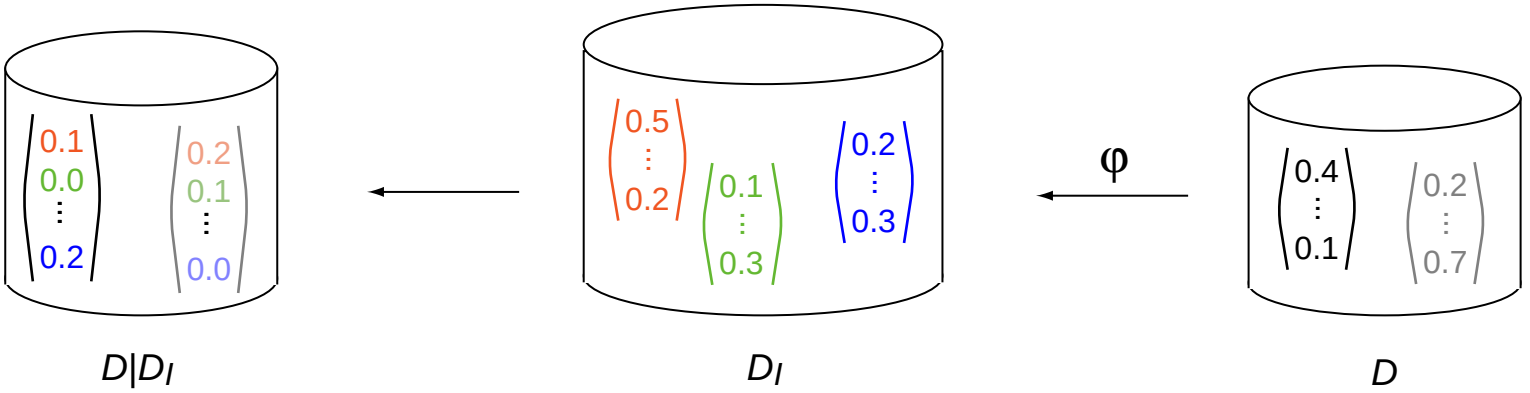
Formalization of CL-ESA

ESA

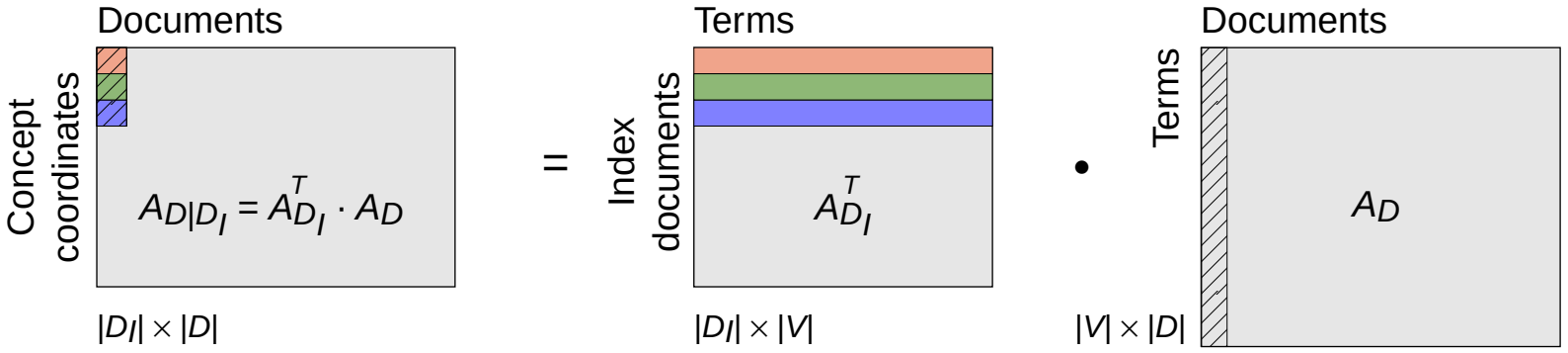


Formalization of CL-ESA

ESA



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



Formalization of CL-ESA

CL-ESA

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

Formalization of CL-ESA

CL-ESA

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

Formalization of CL-ESA

CL-ESA

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

Outlook

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