Karlsruhe Institute of Technology

Combining Concept Based and Text Based Indexes for CLIR

CLEF09: Ad-hoc (TEL) Session, Corfu, Greece

Institute AIFB – University of Karlsruhe

<u>Philipp Sorg</u>
Institute AIFB, Universität Karlsruhe
sorg@kit.edu

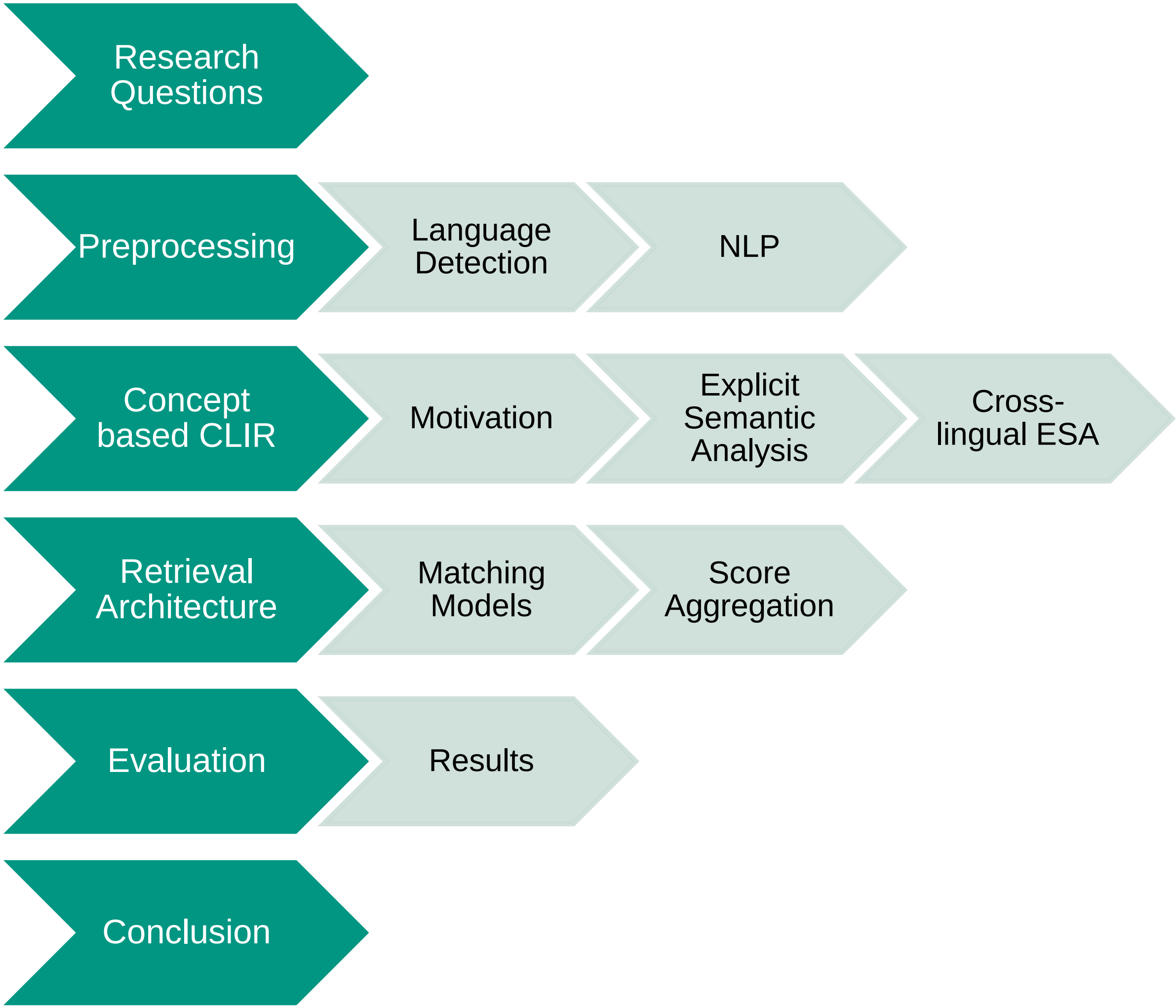
Philipp Cimiano
Web Information Systems Group, Delft University of Technology
p.cimiano@tudelft.nl

Research Questions

- Can multi-lingual information be used to improve retrieval on the TEL dataset?
  - Queries in different languages
  - Documents in different languages
  - Fields of documents in different languages

- Can text based (= Machine Translation based) retrieval be combined with concept based retrieval?
  - Representation of documents in concept space
    - Explicit Semantic Analysis (ESA)
  - Score aggregation problem

Agenda



Research Questions

Preprocessing
- Language Detection
- NLP

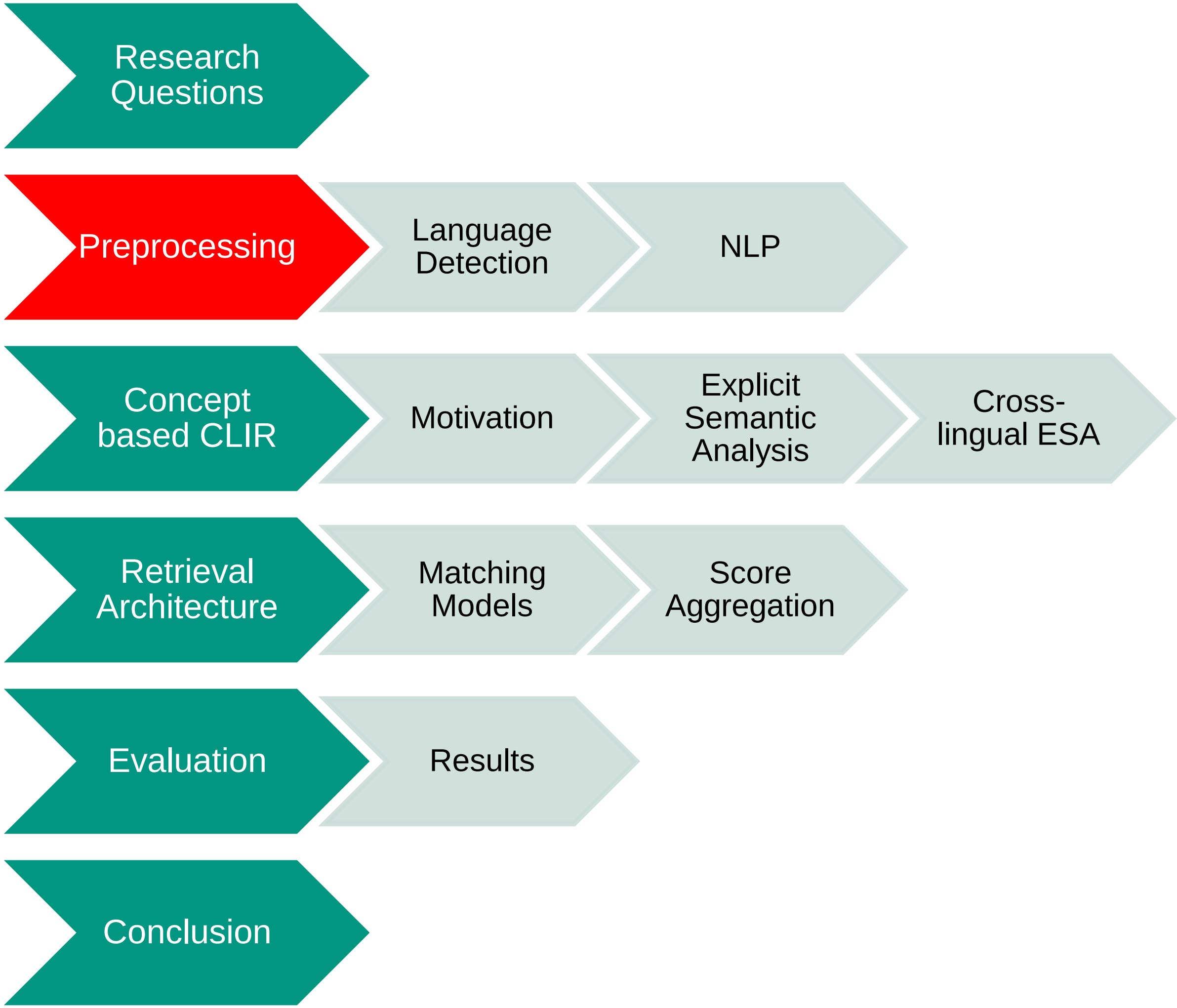
Concept based CLIR
- Motivation
- Explicit Semantic Analysis
- Cross-lingual ESA

Retrieval Architecture
- Matching Models
- Score Aggregation

Evaluation
- Results

Conclusion

Agenda



- Research Questions
- Preprocessing
  - Language Detection
  - NLP
- Concept based CLIR
  - Motivation
  - Explicit Semantic Analysis
  - Cross-lingual ESA
- Retrieval Architecture
  - Matching Models
  - Score Aggregation
- Evaluation
  - Results
- Conclusion

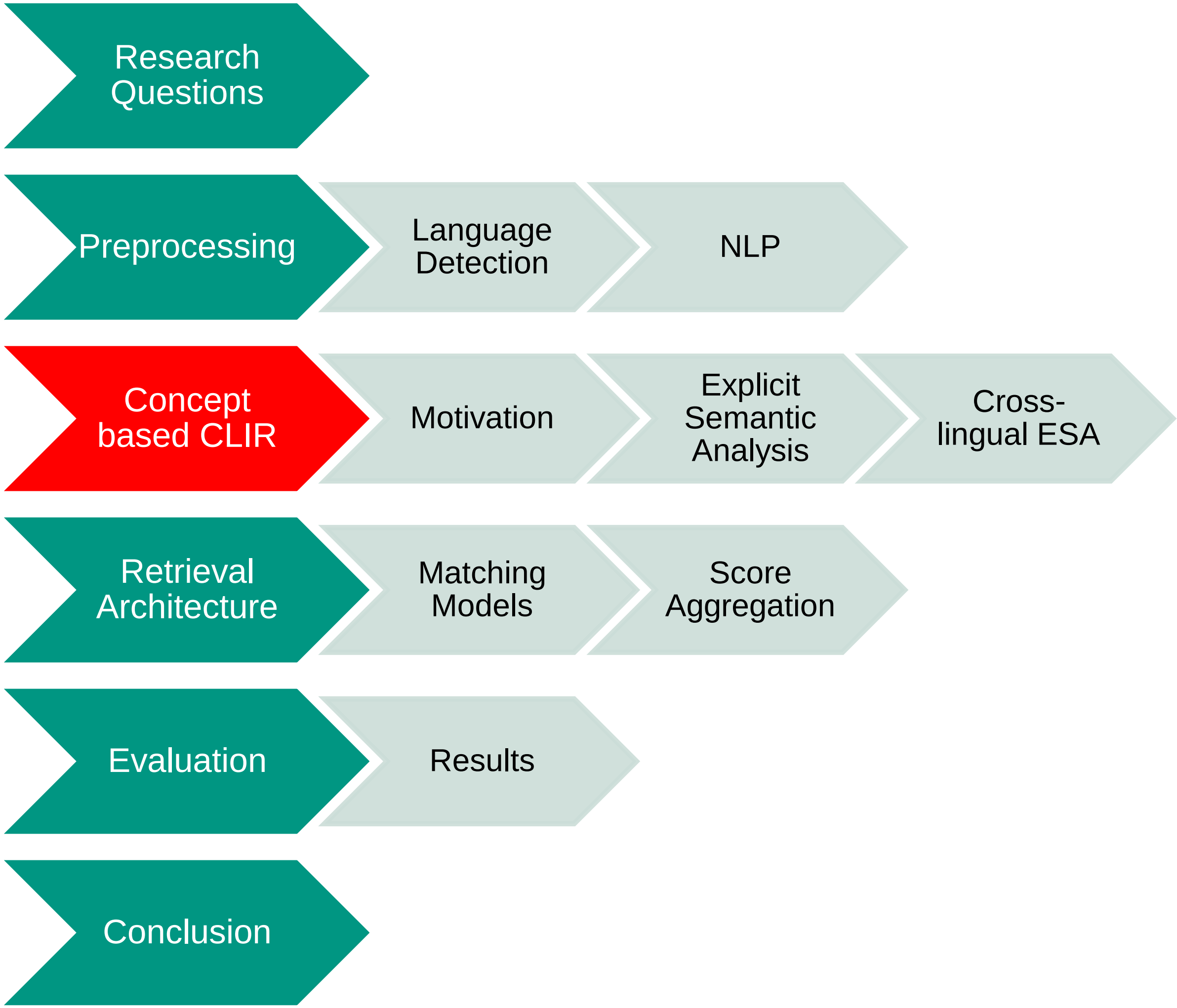
Preprocessing of Dataset

- Selection of content fields
  - Title, subject, alternative, abstract

- Language Detection
  - Character n-gram model for language detection
    - Ling Pipe Identification Tool
  - Each field is classified
    - Based on language tag and language detection
    - Results in documents with multi-lingual fields

- NLP
  - Stemming in all languages supported by Snowball stemmer
  - Language specific stopword removal

Agenda



Research Questions

Preprocessing – Language Detection – NLP

Concept based CLIR – Motivation – Explicit Semantic Analysis – Cross-lingual ESA

Retrieval Architecture – Matching Models – Score Aggregation

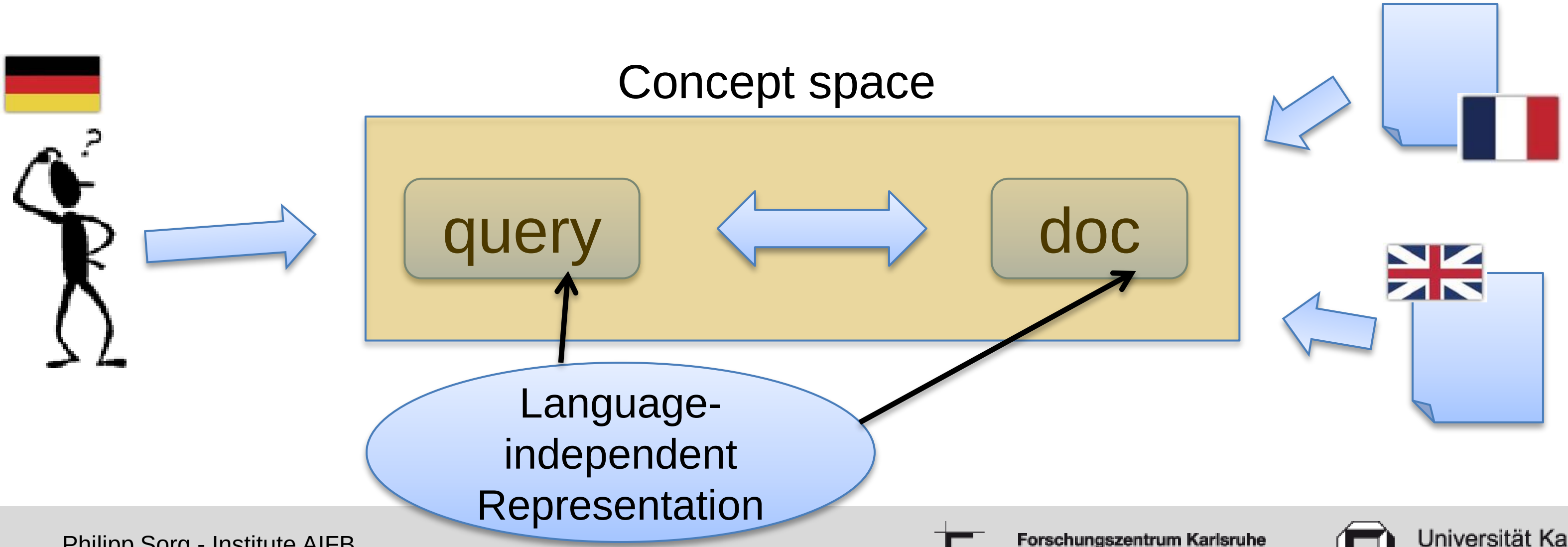
Evaluation – Results

Conclusion

Motivation of Concept Based CLIR

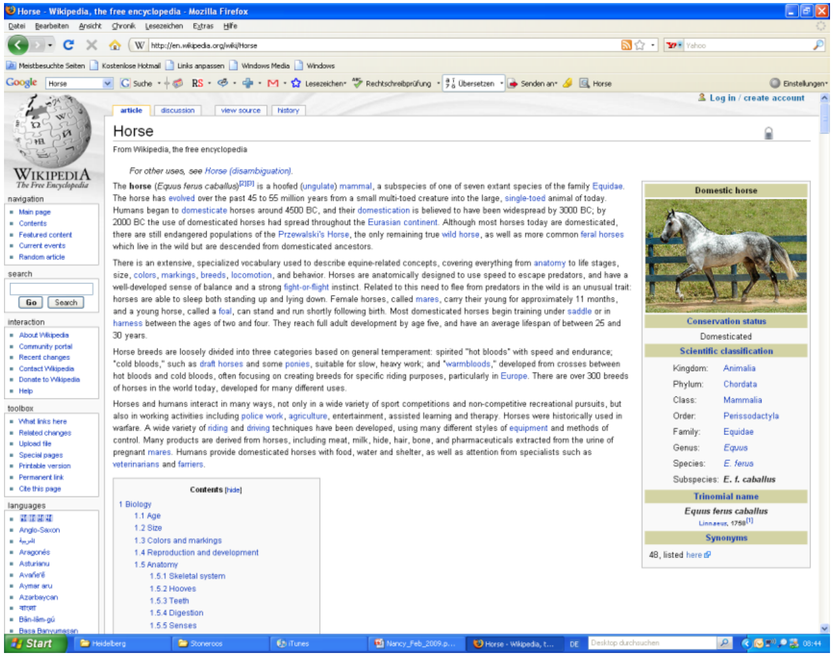
- Traditional approach to Multi-lingual IA
  - Translation of queries or documents
  - Problems
    - MT is not available for many language pairs
    - Propagation of error, inherits all problems of mono-lingual retrieval

- Alternative approach:



Explicit Concept Model

- Idea: Use Web 2.0 resources to define concepts
  - Pragmatic definition of concepts
    - Wikipedia articles, tagged web sites, products, …
  - Cover a broad range of topics and languages
  - Freely available
- Example
  - Wikipedia articles as concepts
- We use Explicit Semantic Analysis (Cross-lingual ESA)
  - Gabrilovich and Markovitch IJCAI 2007
  - Potthast et al. ECIR 2008, Sorg and Cimiano CLEF 2008



Horse

From Wikipedia, the free encyclopedia

For other uses, see Horse (disambiguation).

The **horse** (*Equus ferus caballus*)[2][3] is a hoofed (ungulate) mammal, a subspecies of one of seven extant species of the family Equidae. The horse has evolved over the past 45 to 55 million years from a small multi-toed creature into the large, single-toed animal of today. Humans began to domesticate horses around 4500 BC, and their domestication is believed to have been widespread by 3000 BC; by 2000 BC the use of domesticated horses had spread throughout the Eurasian continent. Although most horses today are domesticated, there are still endangered populations of the Przewalski's Horse, the only remaining true wild horse, as well as more common feral horses which live in the wild but are descended from domesticated ancestors.

There is an extensive, specialized vocabulary used to describe equine-related concepts, covering everything from anatomy to life stages, size, colors, markings, breeds, locomotion, and behavior. Horses are anatomically designed to use speed to escape predators, and have a well-developed sense of balance and a strong fight-or-flight instinct. Related to this need to flee from predators in the wild is an unusual trait: horses are able to sleep both standing up and lying down. Female horses, called mares, carry their young for approximately 11 months, and a young horse, called a foal, can stand and run shortly following birth. Most domesticated horses begin training under saddle or in harness between the ages of two and four. They reach full adult development by age five, and have an average lifespan of between 25 and 30 years.

Horse breeds are loosely divided into three categories based on general temperament: spirited "hot bloods" with speed and endurance; "cold bloods," such as draft horses and some ponies, suitable for slow, heavy work; and "warmbloods," developed from crosses between hot bloods and cold bloods, often focusing on creating breeds for specific riding purposes, particularly in Europe. There are over 300 breeds of horses in the world today, developed for many different uses.

Horses and humans interact in many ways, not only in a wide variety of sport competitions and non-competitive recreational pursuits, but also in working activities including police work, agriculture, entertainment, assisted learning and therapy. Horses were historically used in warfare. A wide variety of riding and driving techniques have been developed, using many different styles of equipment and methods of control. Many products are derived from horses, including meat, milk, hide, hair, bone, and pharmaceuticals extracted from the urine of pregnant mares. Humans provide domesticated horses with food, water and shelter, as well as attention from specialists such as veterinarians and farriers.

Contents [hide]

1 Biology
1.1 Age
1.2 Size
1.3 Colors and markings
1.4 Reproduction and development
1.5 Anatomy
1.5.1 Skeletal system
1.5.2 Hooves
1.5.3 Teeth
1.5.4 Digestion
1.5.5 Senses

Domestic horse

Conservation status

Domesticated

Scientific classification

Kingdom:	Animalia
Phylum:	Chordata
Class:	Mammalia
Order:	Perissodactyla
Family:	Equidae
Genus:	*Equus*
Species:	*E. ferus*
Subspecies:	***E. f. caballus***

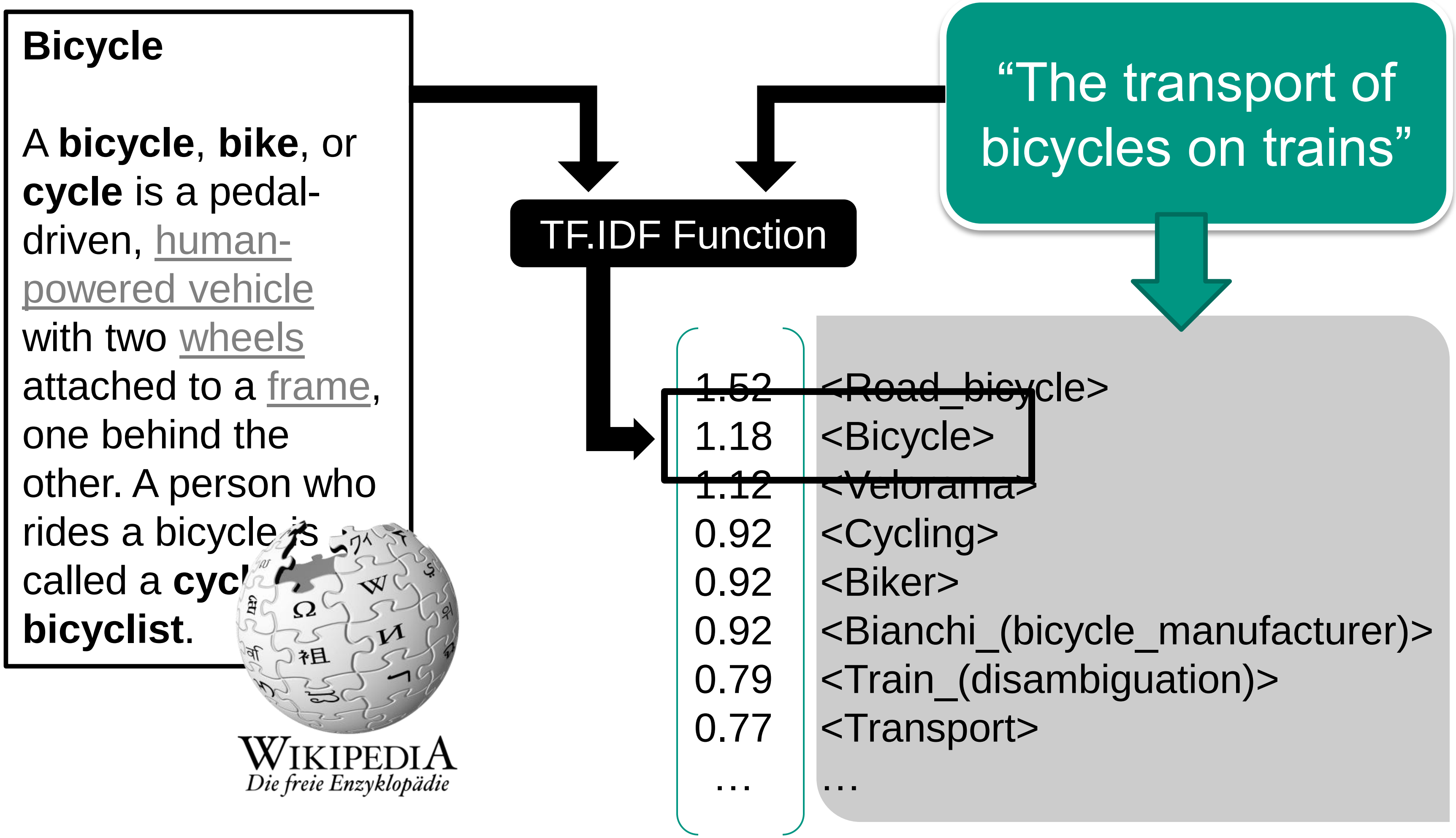
Trinomial name

Equus ferus caballus
Linnaeus, 1758[1]

Synonyms

48, listed here

Idea of ESA



Bicycle

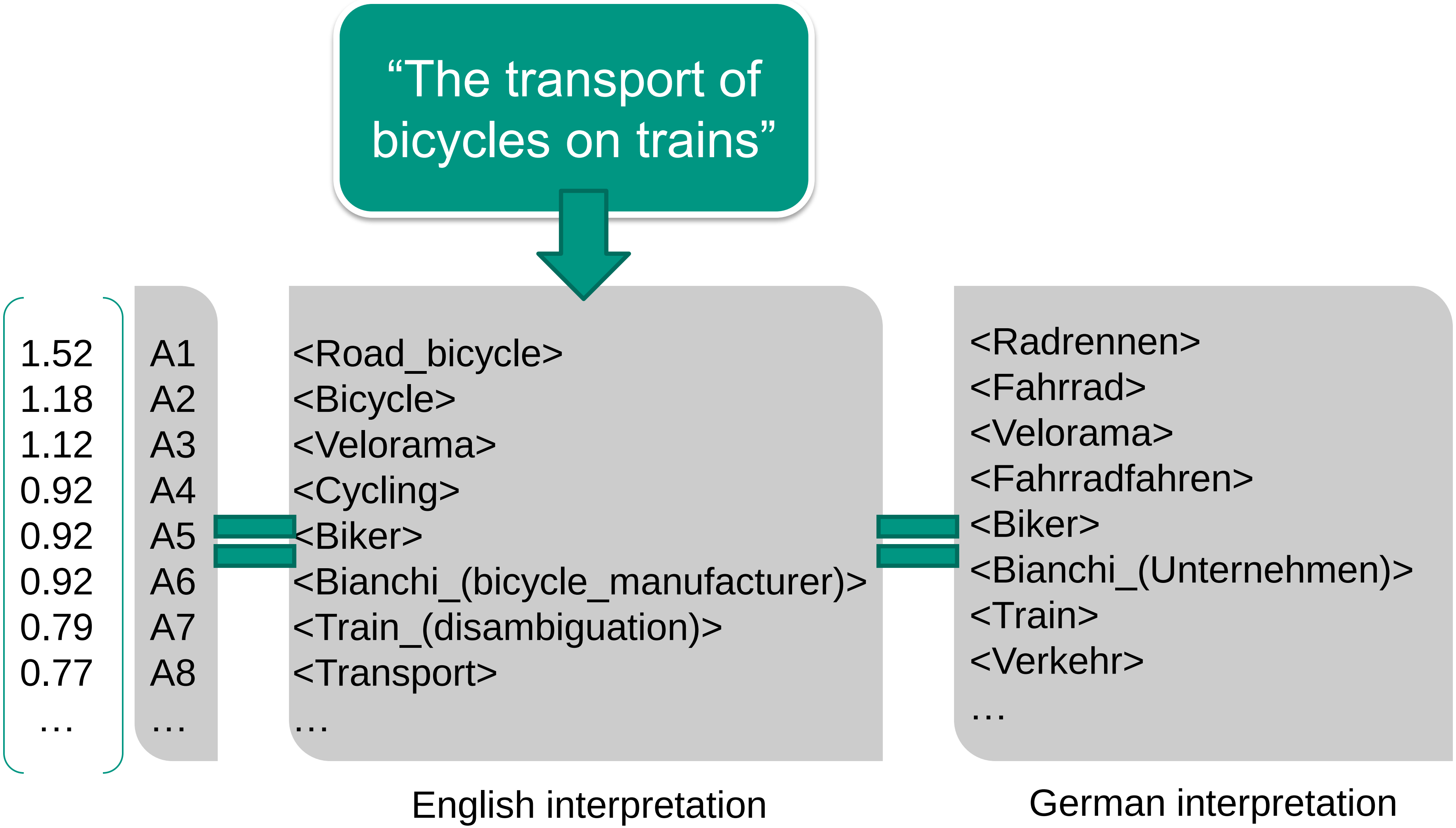
A **bicycle**, **bike**, or **cycle** is a pedal-driven, human-powered vehicle with two wheels attached to a frame, one behind the other. A person who rides a bicycle is called a **cycl**ist or **bicyclist**.

"The transport of bicycles on trains"

TF.IDF Function

1.52	<Road_bicycle>
1.18	<Bicycle>
1.12	<Velorama>
0.92	<Cycling>
0.92	<Biker>
0.92	<Bianchi_(bicycle_manufacturer)>
0.79	<Train_(disambiguation)>
0.77	<Transport>
…	…

Example Cross-lingual ESA Concept Vector



“The transport of bicycles on trains”

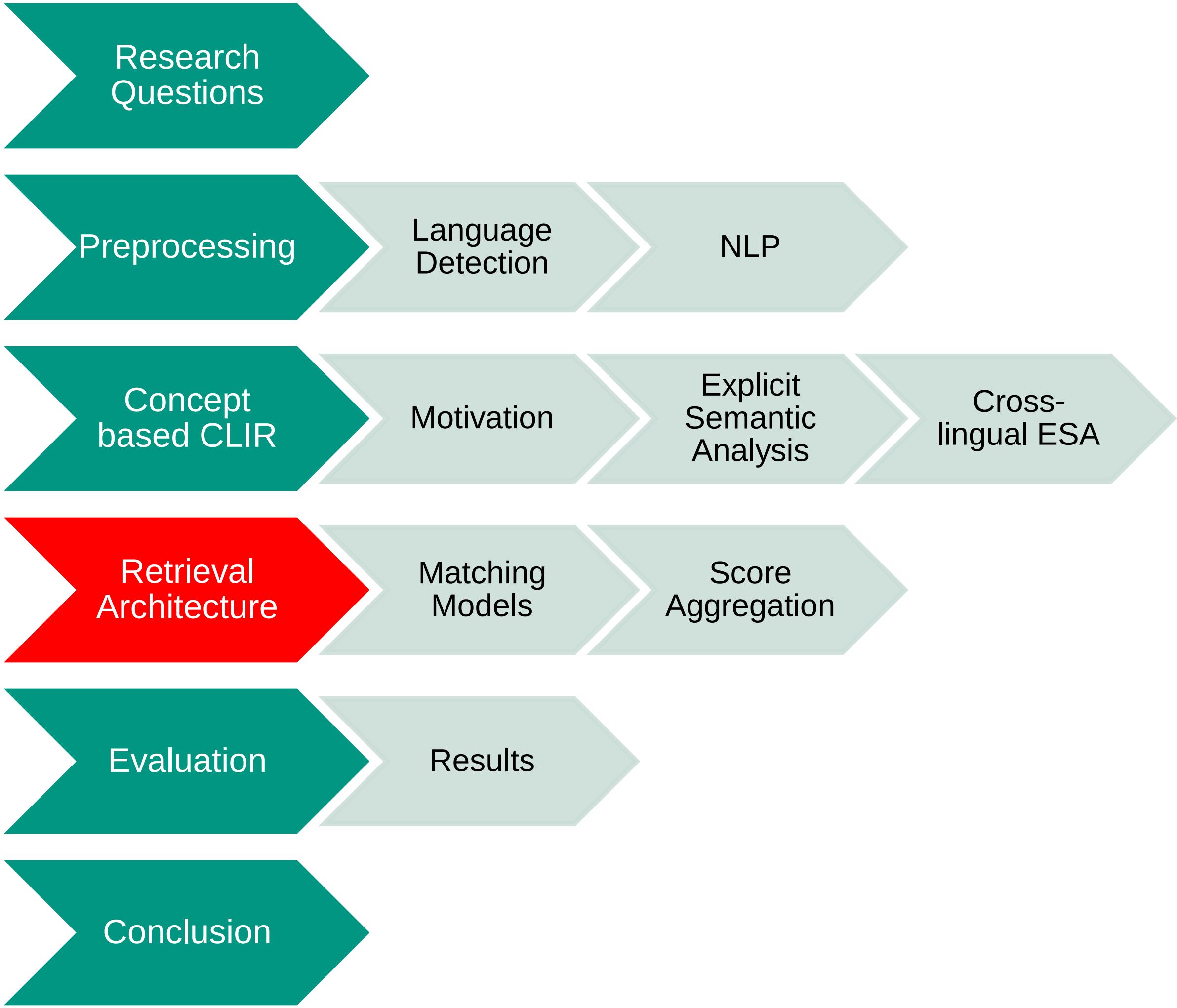
1.52 A1 <Road_bicycle>
1.18 A2 <Bicycle>
1.12 A3 <Velorama>
0.92 A4 <Cycling>
0.92 A5 <Biker>
0.92 A6 <Bianchi_(bicycle_manufacturer)>
0.79 A7 <Train_(disambiguation)>
0.77 A8 <Transport>
… … …

<Radrennen>
<Fahrrad>
<Velorama>
<Fahrradfahren>
<Biker>
<Bianchi_(Unternehmen)>
<Train>
<Verkehr>
…

English interpretation

German interpretation

Agenda



Research Questions

Preprocessing – Language Detection – NLP

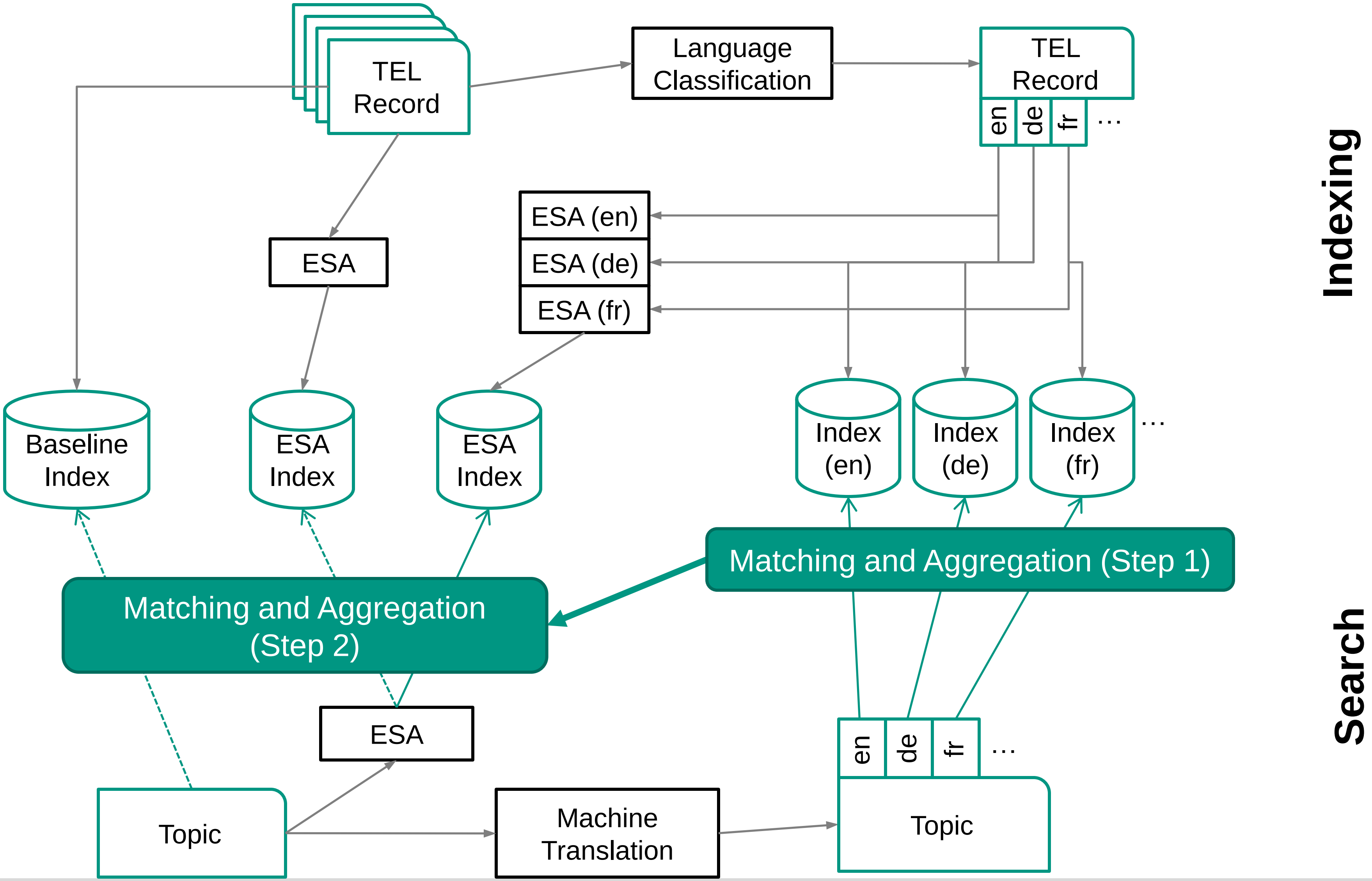
Concept based CLIR – Motivation – Explicit Semantic Analysis – Cross-lingual ESA

Retrieval Architecture – Matching Models – Score Aggregation

Evaluation – Results

Conclusion

Retrieval Architecture



Matching and Aggregation (Step 1)

- Optimization of matching model
  - Using CLEF2008 topics and relevance assessments
  - Models provided by the Terrier framework
    - BL: DLH13, ONB: LemurTF_IDF, BNF: BB2

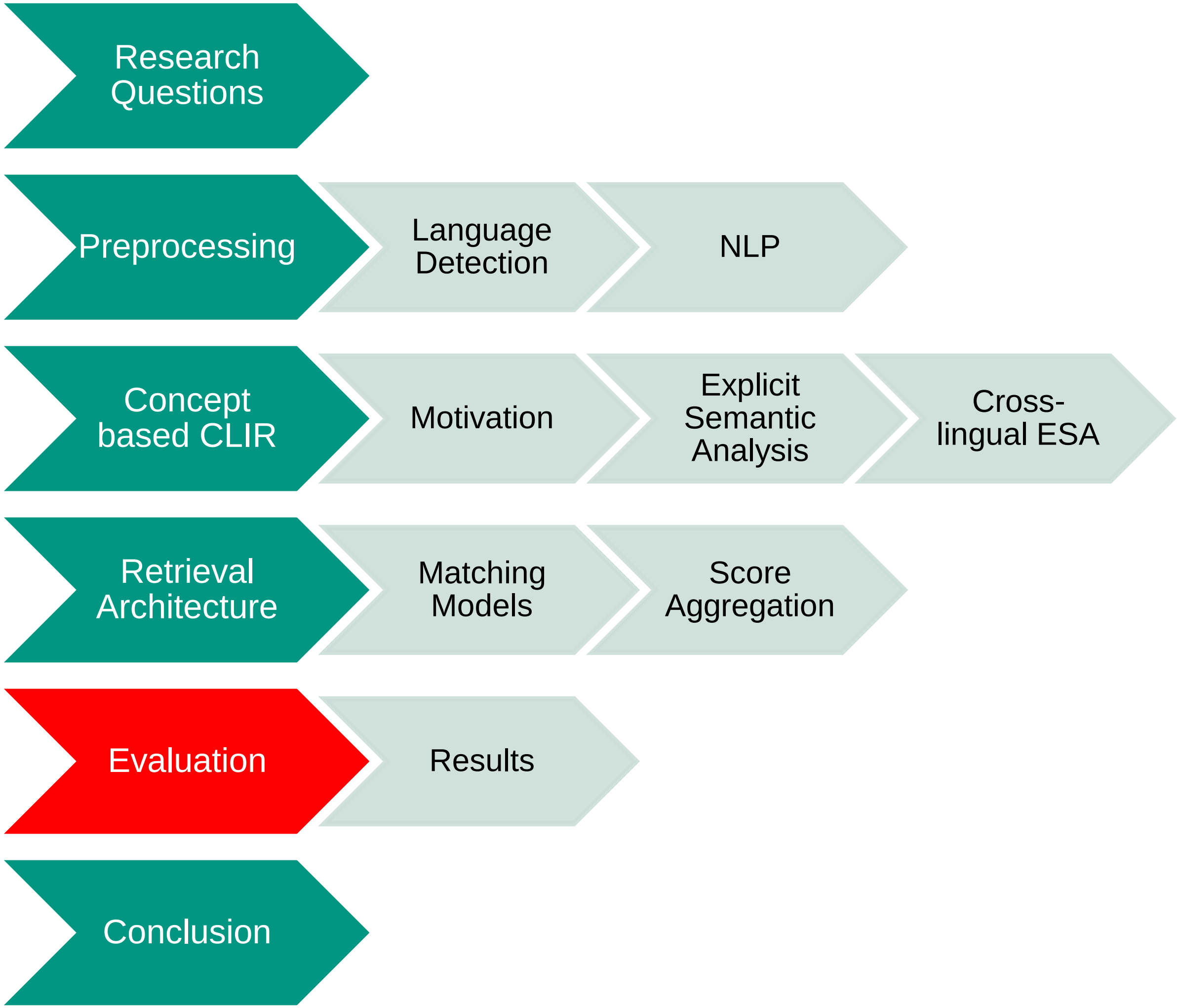
- Linear aggregation of scores
  - Each document has a score for each index (=language)
  - Different normalization functions
    - Based on maximal score in each ranking
    - Based on the number of retrieved documents of each ranking
    - Based on a priori weights
      - Language distribution of text in corpus

$$score(t,d) := \sum_{r \in R} \delta(r)\ score_r(t,d)$$

Matching and Aggregation (Step 2)

- ESA retrieval using cosine similarity
  - Implementation based on inverted concept index
- Linear aggregation of concept based scores and text based scores
  - Using the aggregated score from text based retrieval (Step 1)
  - Weight factor to modify influence of concept based retrieval
    - Optimized on CLEF2008 topics
- Evaluation measures
  - MAP: Mean Average Precision
  - P@10: Precision at cutoff level of 10 documents

Agenda



Research Questions

Preprocessing
Language Detection
NLP

Concept based CLIR
Motivation
Explicit Semantic Analysis
Cross-lingual ESA

Retrieval Architecture
Matching Models
Score Aggregation

Evaluation
Results

Conclusion

Evaluation

Topic lang.	Retrieval Method	BL		ONB		BNF	
		MAP	P@10	MAP	P@10	MAP	P@10
En	Baseline (single index)	35	51	16	26	25	39
	Multiple Indexes	33	50	15	24	22	34
	Concept + Baseline	35	52	**17**	27	25	39
De	Baseline (single index)	33	49	23	35	24	35
	Multiple Indexes	31	48	23	34	22	32
	Concept + Baseline	33	49	**24**	35	24	36
Fr	Baseline (single index)	31	48	15	22	27	38
	Multiple Indexes	29	45	14	20	25	35
	Concept + Baseline	32	**51**	15	22	27	37

Evaluation

Topic lang.	Retrieval Method	BL		ONB		BNF	
		MAP	P@10	MAP	P@10	MAP	P@10
En	Baseline (single index)	35	51	16	26	25	39
	Multiple Indexes	33	50	15	24	22	34
	Concept + Baseline	35	52	**17**	27	25	39
De	Baseline (single index)	33	49	23	35	24	35
	Multiple Indexes	31	48	23	34	22	32
	Concept + Baseline	33	49	**24**	35	24	36
Fr	Baseline (single index)	31	48	15	22	27	38
	Multiple Indexes	29	45	14	20	25	35
	Concept + Baseline	32	**51**	15	22	27	37

Evaluation

Topic lang.	Retrieval Method	BL		ONB		BNF	
		MAP	P@10	MAP	P@10	MAP	P@10
En	Baseline (single index)	35	51	16	26	25	39
	Multiple Indexes	33	50	15	24	22	34
	Concept + Baseline	35	52	**17**	27	25	39
De	Baseline (single index)	33	49	23	35	24	35
	Multiple Indexes	31	48	23	34	22	32
	Concept + Baseline	33	49	**24**	35	24	36
Fr	Baseline (single index)	31	48	15	22	27	38
	Multiple Indexes	29	45	14	20	25	35
	Concept + Baseline	32	**51**	15	22	27	37

Evaluation

Topic lang.	Retrieval Method	BL		ONB		BNF	
		MAP	P@10	MAP	P@10	MAP	P@10
En	Baseline (single index)	35	51	16	26	25	39
	Multiple Indexes	33	50	15	24	22	34
	Concept + Baseline	35	52	**17**	27	25	39
De	Baseline (single index)	33	49	23	35	24	35
	Multiple Indexes	31	48	23	34	22	32
	Concept + Baseline	33	49	**24**	35	24	36
Fr	Baseline (single index)	31	48	15	22	27	38
	Multiple Indexes	29	45	14	20	25	35
	Concept + Baseline	32	**51**	15	22	27	37

Conclusion

- Baseline is very strong
- Can multi-lingual information be used to improve retrieval on the TEL dataset?
  - Use of multi-lingual indexes based on language detection did not improve retrieval
    - Problem of score aggregation
    - Linear aggregation model with (simple) normalization is not working
- Can text based (= Machine Translation based) retrieval be combined with concept based retrieval?
  - Combination of concept and text based indexes yields only small improvements
    - We could not reconstruct the large improvements reported on mono-lingual collections
    - Not enough context in short TEL records for concept mapping?

Thank you!

- Questions?

- Joint work with
  - Philipp Cimiano (Universität Bielefeld)
  - Marlon Braun, David Nicolay (Universität Karlsruhe)

- Acknowledgments
  - Multipla Project
    DFG grant 38457858

MULTIPLA Multi-Ontology Learning