FONDAZIONE BRUNO KESSLER

Using Semantic and Domain-based Information in CLIR Systems

Alessio Bosca[2], Matteo Casu[2], Chiara Di Francescomarino[1], **Mauro Dragoni[1]**

(1) Fondazione Bruno Kessler (FBK), Shape and Evolve Living Knowledge Unit (SHELL)

(2) CELI s.r.l.

https://shell.fbk.eu/index.php/Mauro_Dragoni - dragoni@fbk.eu

11th Extended Semantic Web Conference 2014 – May, 27th 2014

FONDAZIONE BRUNO KESSLER

Background – CLIR: 3 Scenarios…

- The document collection is monolingual, but users can formulate queries in more than one language.
- The document collection contains documents in multiple languages and users can query the entire collection in one or more languages.
- The document collection contains documents with mixed-language content and users can query the entire collection in one or more languages.

FONDAZIONE BRUNO KESSLER

Background – … and 2 strategies

- Model dependent
  - Translation and retrieval are integrated in an uniform framework
- Model independent
  - Translation and retrieval are treated as separated processes

FONDAZIONE
BRUNO KESSLER

Background - Challenges

- Out-of-Vocabulary issue

  → improve the corpora used for training the machine translation model.

  → usage of domain information for increasing the coverage of the dictionaries.

- Usage of semantic artifacts for structuring the representation of (multilingual) documents.

Background - Challenges

- Out-of-Vocabulary issue

  → improve the corpora used for training the machine translation model.

  → usage of domain information for increasing the coverage of the dictionaries.

- Usage of semantic artifacts for structuring the representation of (multilingual) documents.

GOAL

to integrate domain-specific semantic knowledge within a CLIR system and evaluate their effectiveness

FONDAZIONE
BRUNO KESSLER

Our Scenario

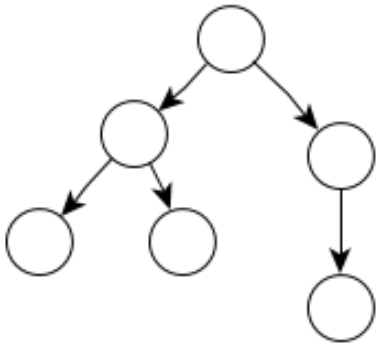
- Use case: the agricultural domain
- Knowledge resources: Agrovoc and Organic.Lingua ontologies
- 3 components used in the proposed approach:
  - Annotator
  - Indexer
  - Retriever

FONDAZIONE
BRUNO KESSLER

Annotation Process – Step 1

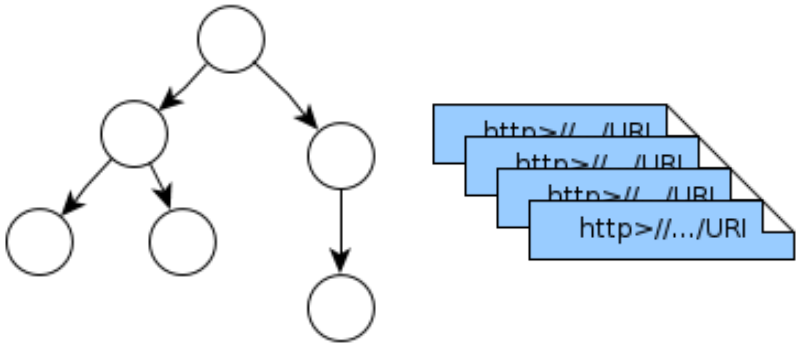
FONDAZIONE
BRUNO KESSLER

Annotation Process – Step 1

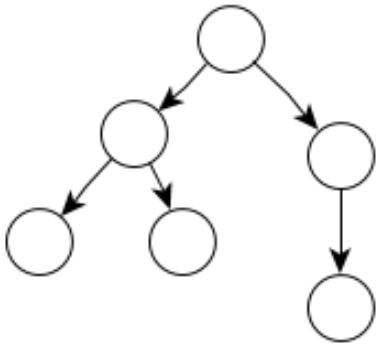


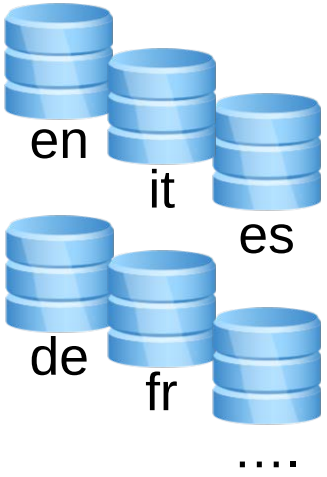
FONDAZIONE
BRUNO KESSLER

Annotation Process – Step 1

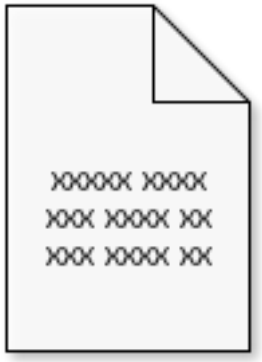


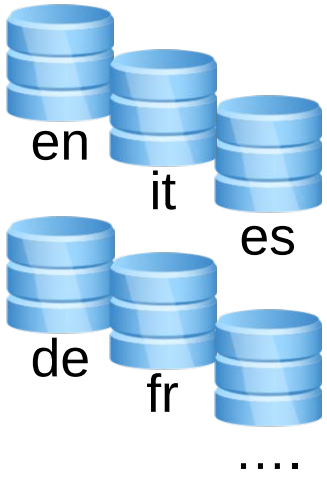
Annotation Process – Step 1





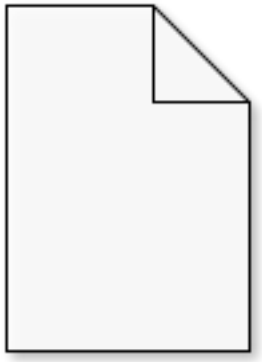
Annotation Process – Step 2

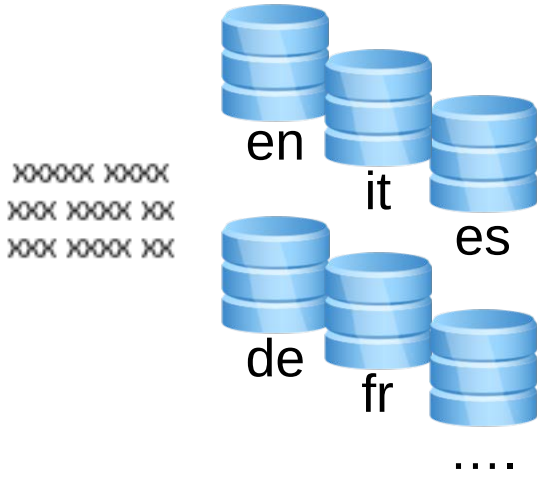




- Document content is used as query.
- Between the candidate results, only “exact matches” are considered.

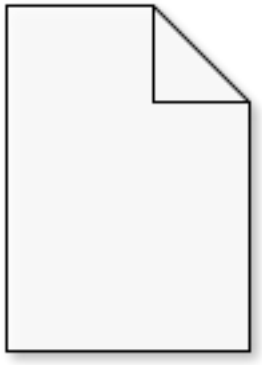
Annotation Process – Step 2

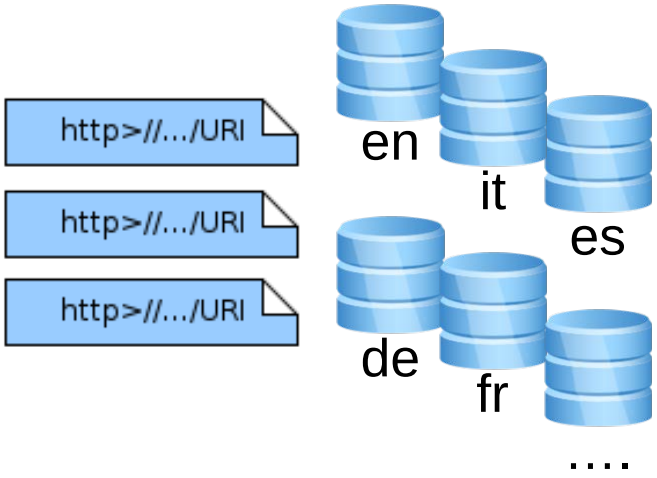




- Document content is used as query.
- Between the candidate results, only “exact matches” are considered.

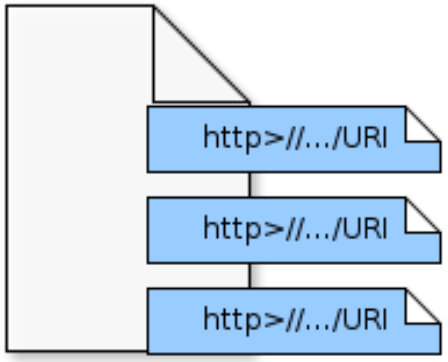
Annotation Process – Step 2

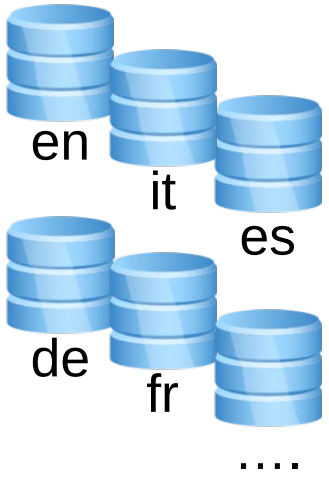




- Document content is used as query.
- Between the candidate results, only “exact matches” are considered.

Annotation Process – Step 2





- Document content is used as query.
- Between the candidate results, only “exact matches” are considered.

FONDAZIONE BRUNO KESSLER

Approach – Annotation Stats

Domain Ontology	Number of Concepts	Manual Annotations	Automatic Annotations
Agrovoc (AV)	32061	0	133596 (5834 distinct concepts used)
Organic.Lingua (OL)	291	27871 (264 distinct concepts used)	16434 (208 distinct concepts used)

FONDAZIONE
BRUNO KESSLER

Approach - Index

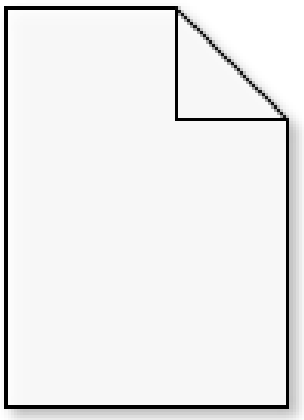
- Given a document:
  - Text and annotations are extracted.
  - The context of each concept is retrieved from the ontologies.
  - Each contextual concepts are indexed with a weight proportional w.r.t. their semantic distance from the semantic annotation.

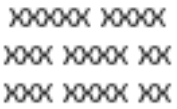
.

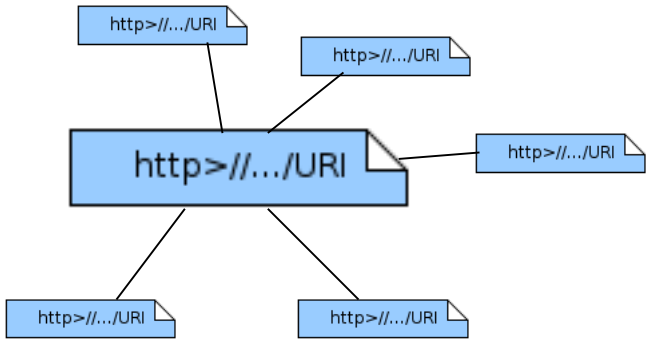
Approach - Index

- Given a document:
  - Text and annotations are extracted.
  - The context of each concept is retrieved from the ontologies.
  - Each contextual concepts are indexed with a weight proportional w.r.t. their semantic distance from the semantic annotation.

- Structure of each index record:







FONDAZIONE BRUNO KESSLER

Approach - Retriever

- Three retrieval configurations available:
  - Only translations: query terms are translated by using machine translation services.
  - Semantic expansion by exploiting the domain ontology: query terms are matched with ontology concepts; if an exact match exists, query is expanded by using the URI of the concept and the URIs of the contextual ones.
  - Ontology matching only: terms not having an exact match with ontology concepts are discarded.

FONDAZIONE BRUNO KESSLER

Evaluation - Setup

- Collection of 13,000 multilingual documents.
- 48 queries originally provided in English and manually translated in 12 languages under the supervision of both domain and language experts.
- Gold standard manually built by the domain experts.
- MAP, Prec@5, Prec@10, Prec@20, Recall have been used.

FONDAZIONE
BRUNO KESSLER

Results - 1

	Avg. MAP	Prec@5	Prec@10	Prec@20	Avg. Rec.
BASELINE	**0.554**	**0.617**	**0.545**	**0.465**	**0.920**

FONDAZIONE BRUNO KESSLER

Results - 1

	Avg. MAP	Prec@5	Prec@10	Prec@20	Avg. Rec.
BASELINE	**0.554**	**0.617**	**0.545**	**0.465**	**0.920**
Auto: AV	3.24%	3.11%	5.04%	3.81%	2.52%
Auto: OL	2.31%	1.91%	2.88%	2.98%	0.77%
Auto: AV+OL	3.13%	2.95%	4.63%	3.86%	2.53%

FONDAZIONE BRUNO KESSLER

Results - 1

	Avg. MAP	Prec@5	Prec@10	Prec@20	Avg. Rec.
BASELINE	**0.554**	**0.617**	**0.545**	**0.465**	**0.920**
Auto: AV	3.24%	3.11%	5.04%	3.81%	2.52%
Auto: OL	2.31%	1.91%	2.88%	2.98%	0.77%
Auto: AV+OL	3.13%	2.95%	4.63%	3.86%	2.53%
Auto+Man: OL	1.65%	3.40%	3.95%	4.48%	1.37%
Auto+Man: AV+OL	4.38%	5.96%	7.18%	6.07%	2.97%

FONDAZIONE BRUNO KESSLER

Results - 1

	Avg. MAP	Prec@5	Prec@10	Prec@20	Avg. Rec.
BASELINE	**0.554**	**0.617**	**0.545**	**0.465**	**0.920**
Auto: AV	3.24%	3.11%	5.04%	3.81%	2.52%
Auto: OL	2.31%	1.91%	2.88%	2.98%	0.77%
Auto: AV+OL	3.13%	2.95%	4.63%	3.86%	2.53%
Auto+Man: OL	1.65%	3.40%	3.95%	4.48%	1.37%
Auto+Man: AV+OL	4.38%	5.96%	7.18%	6.07%	2.97%
Auto+Man*2: OL	1.00%	3.30%	4.02%	3.27%	1.36%
Auto+Man*2: AV+OL	3.29%	4.86%	6.73%	6.03%	2.97%

FONDAZIONE BRUNO KESSLER

Results - 2

	Query Cov.	Avg. MAP	Prec@5	Prec@10	Prec@20	Avg. Rec.
AV	39.3 (9 langs)	0.137	0.189	0.191	0.179	0.552
OL	15.7 (10 langs)	0.260	0.359	0.319	0.322	0.635
AV + OL	33.3 (12 langs)	0.173	0.247	0.226	0.221	0.586

FONDAZIONE
BRUNO KESSLER

Conclusions

- The use of domain-specific ontologies lead to an improvement of CLIR systems effectiveness.
- Find the right trade-off between the effort of manually annotating documents and the system effectiveness

Future work:

- Improve the automatic annotation component
- Move to a more complex semantic representation of information in order to answer to more complex query.

References:

- www.organic-edunet.eu: the portal
- www.organic-lingua.eu/en/outcomes/deliverables: the data

FONDAZIONE
BRUNO KESSLER

Mauro Dragoni

https://shell.fbk.eu/index.php/Mauro_Dragoni
dragoni@fbk.eu