From Monolingual to Multilingual Ontologies: The Role of Cross-lingual Ontology Enrichment

Shimaa Ibrahim, Said Fathalla, Hamed Shariat Yazdi, Jens Lehmann, and Hajira Jabeen

SEMANTiCS19 - 15th International Conference on Semantic Systems
Karlsruhe, 9 - 12 September 2019

Outline

- Background
- Motivation
- Objective
- Our previous work
- What is new in the current approach?
- The proposed approach
- Use case
- Evaluation
- Conclusion
- Future work

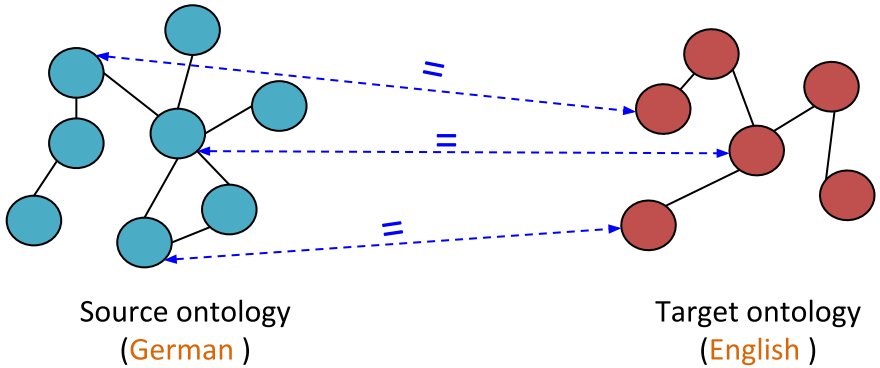
Background

- Multilingual ontology
  - Entities and relations are presented in *several natural languages*.

```
###  http://dbpedia.org/ontology/Bacteria
:Bacteria rdf:type owl:Class ;
          rdfs:subClassOf :Species ;
          rdfs:label "bacteria"@en ,
                     "bacterie"@nl ,
                     "bactérie"@fr ,
                     "bakterium"@de ,
                      prov:wasDerivedFrom
<http://mappings.dbpedia.org/index.php/OntologyClass:Bacteria>
.
```

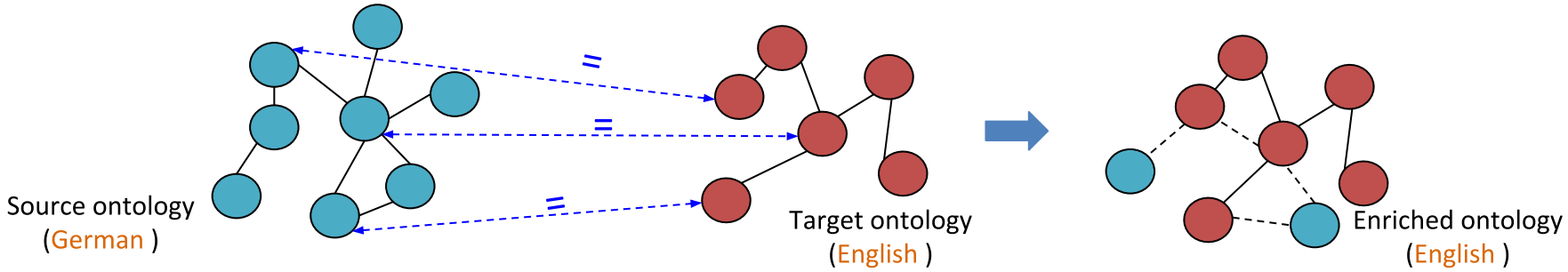
Background

- Cross-lingual ontology matching
  - Match a source ontology to a target ontology *in a different natural language*.



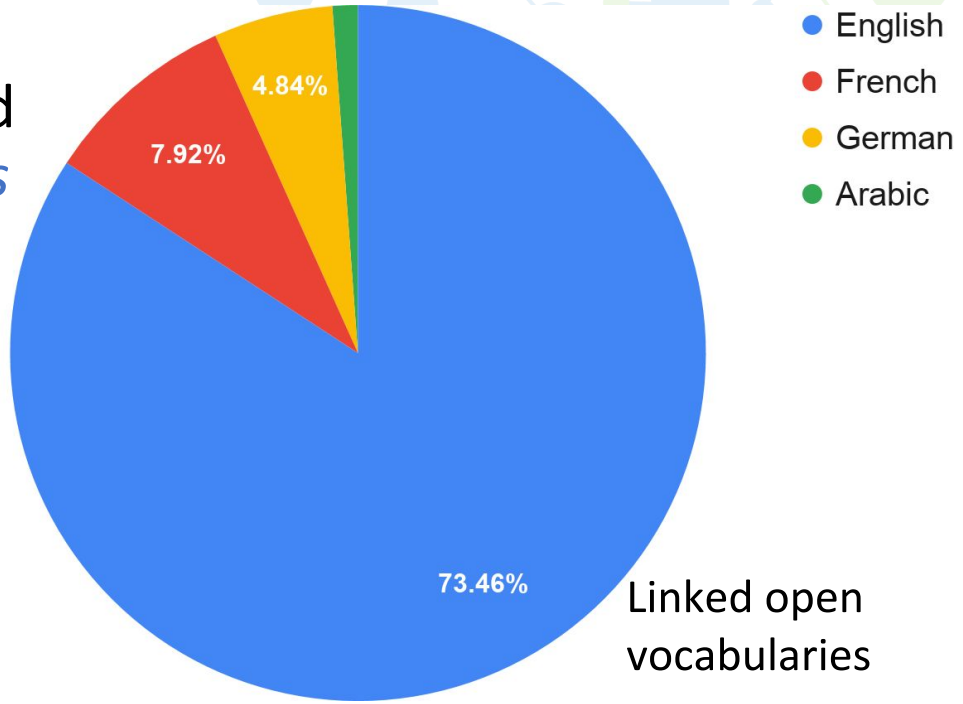
Background

- Cross-lingual ontology enrichment
  - Depends on the cross-lingual matching task,
  - Expand the target ontology with additional information extracted from external resources in other natural languages.



Motivation

- Many ontologies scattered across the web *in various natural languages.*

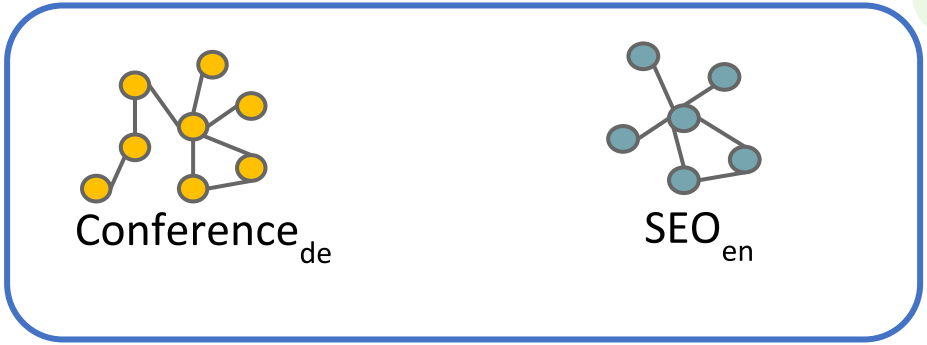


Linked open vocabularies

Motivation

- Manual ontology enrichment is an error-prone and time-consuming task.
- Most of the existing work focus on enriching English ontologies from English sources only (monolingual enrichment).

Motivation Example



- Monolingual ontologies are not easily understandable to speakers of other languages.

Objective

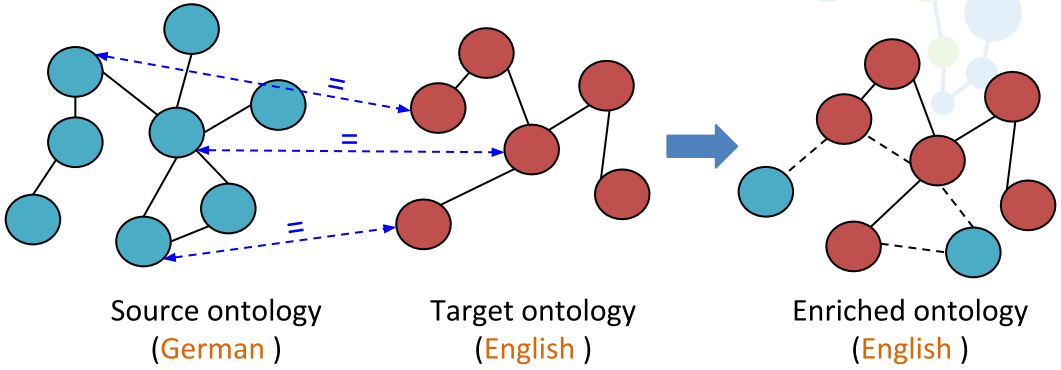
Building multilingual ontologies from the existing monolingual ones in order to *enhance semantic interoperability* and benefit knowledge-based applications.

Research Question

RQ: How can we automatically build multilingual ontologies from monolingual ones?

Our Previous Work

- OECM: A Cross-lingual Approach for Ontology Enrichment - poster @ESWC 2019



What is New in the Current Approach?

- The use of semantic similarity measures.
- Enriching the target ontology by adding new classes in addition to all their related classes in the hierarchy.
- The use of ontologies in non-Indo-European languages (e.g., Arabic), as the source of information.
- Building multilingual ontologies.
- Fully automated approach.

The Proposed Approach: OECM

Problem Formulation

- Given two ontologies **S** and **T**, in two different natural languages L_s and L_t, as RDF triples <s,p,o> $\in C \times R \times (C \cup L)$, Finding the complementary information T_e from **S** to enrich **T:**

$$T_e = S - (S \cap T)$$

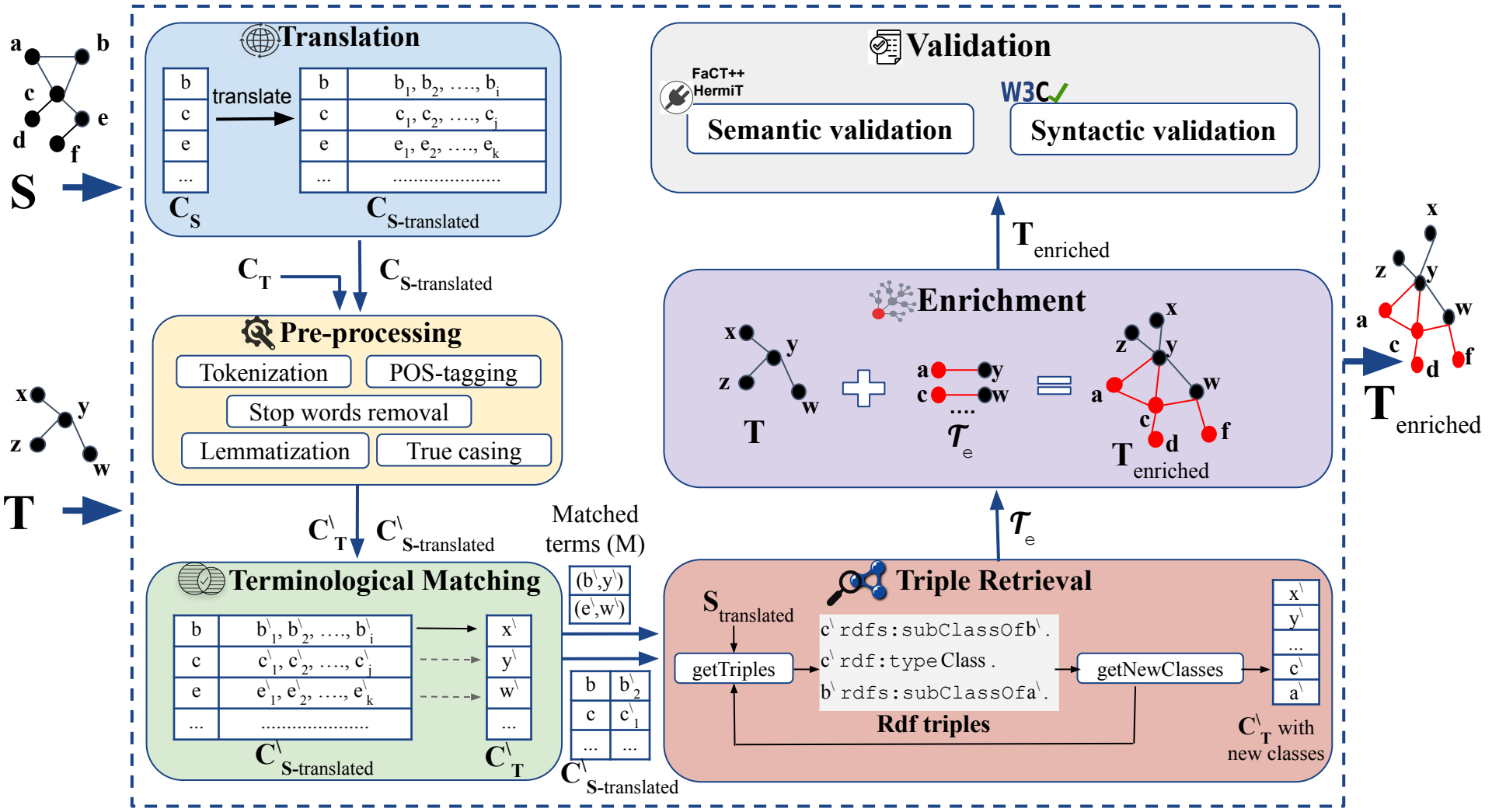
$$T_{enriched} = T_e + T$$

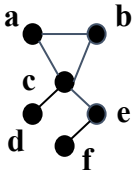
C: the set of ontology domain entities (i.e. classes)

R : the set of relations

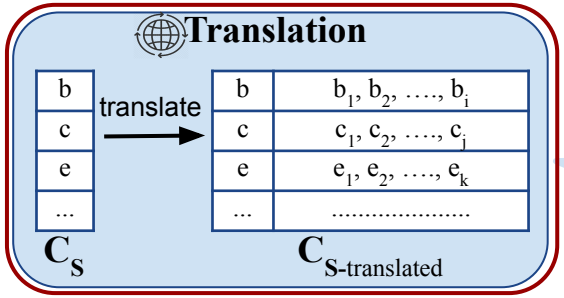
L: the set of literals

OECM Architecture


S
T
Translation
translate
C_S
$C_{S\text{-translated}}$
C_T
Pre-processing
Tokenization
POS-tagging
Stop words removal
Lemmatization
True casing
C'_T
$C'_{S\text{-translated}}$
Terminological Matching
$C'_{S\text{-translated}}$
C'_T
Matched terms (M)
$C'_{S\text{-translated}}$
$S_{translated}$
Triple Retrieval
getTriples
c' rdfs:subClassOf b'.
c' rdf:type Class .
b' rdfs:subClassOf a'.
Rdf triples
getNewClasses
C'_T with new classes
τ_e
Enrichment
T
τ_e
$T_{enriched}$
Validation
FaCT++
HermiT
Semantic validation
W3C
Syntactic validation
$T_{enriched}$
$T_{enriched}$

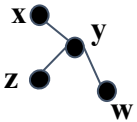


S

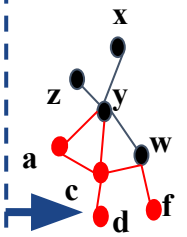


- Translate local names and/or labels of classes of **S** to the language of **T** using Google translator
- Consider all available translations

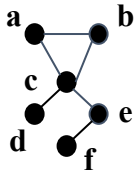
ex: “Thema”in German --> “Subject”, and ”Topic” in English



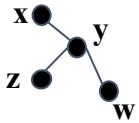
T



$T_{enriched}$



S



T

Translation

	translate		
b	→	b	$b_1, b_2, \ldots, b_i$
c		c	$c_1, c_2, \ldots, c_j$
e		e	$e_1, e_2, \ldots, e_k$
...		...	.....

C_S $C_{S\text{-translated}}$

C_T $C_{S\text{-translated}}$

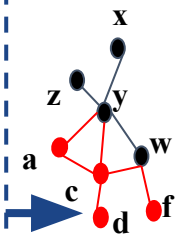
Pre-processing

Tokenization | POS-tagging

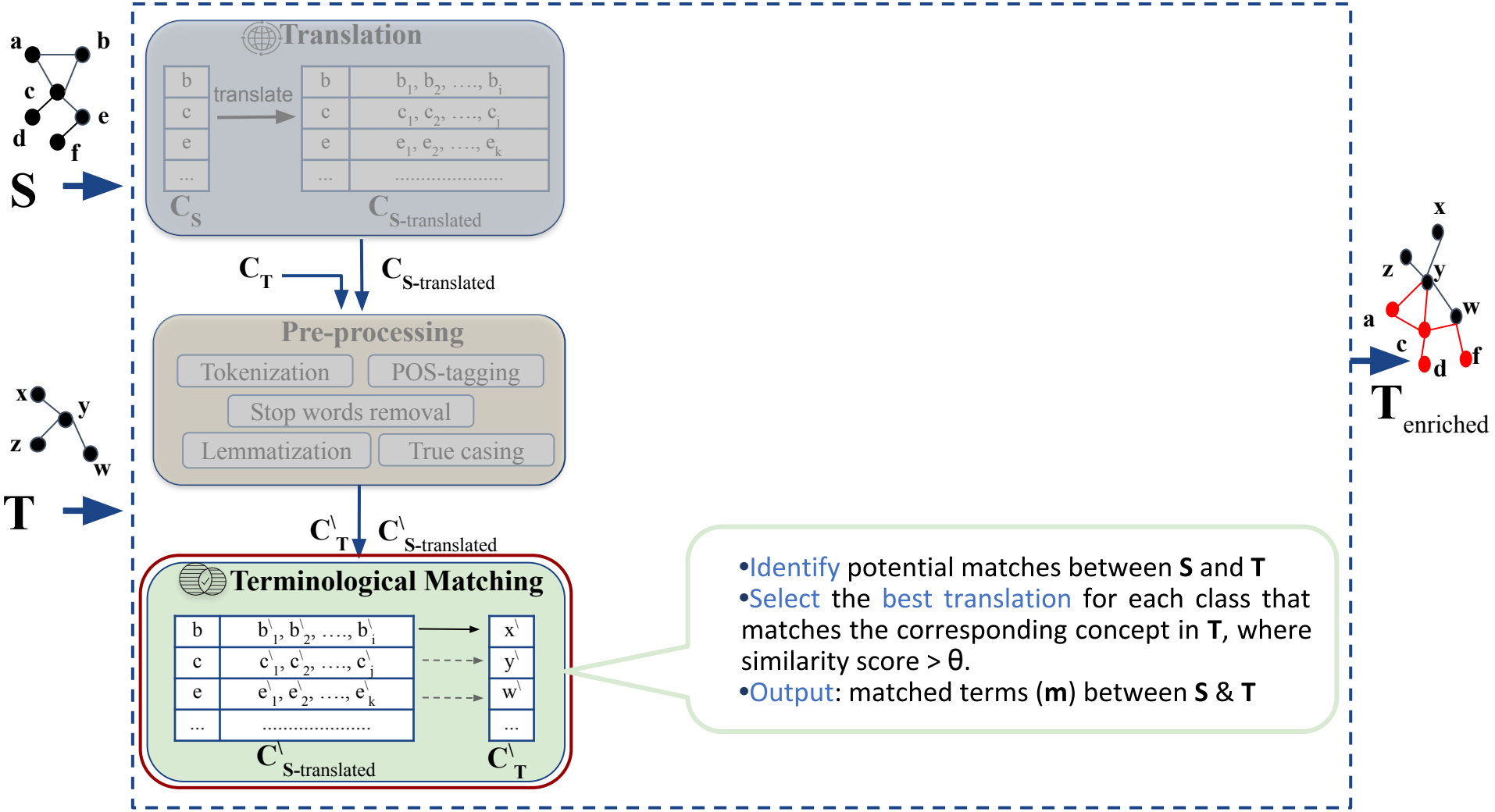
Stop words removal

Lemmatization | True casing

- Process classes of **T** and lists of translations for **S** using NLP techniques



$T_{enriched}$


a
b
c
e
d
f
S
Translation
b
c
e
...
translate
b
c
e
...
$b_1, b_2, \ldots, b_i$
$c_1, c_2, \ldots, c_j$
$e_1, e_2, \ldots, e_k$
C_S
$C_{S\text{-translated}}$
C_T
$C_{S\text{-translated}}$
Pre-processing
Tokenization
POS-tagging
Stop words removal
Lemmatization
True casing
x
y
z
w
T
C'_T
$C'_{S\text{-translated}}$
Terminological Matching
$b'_1, b'_2, \ldots, b'_i$
$c'_1, c'_2, \ldots, c'_j$
$e'_1, e'_2, \ldots, e'_k$
x'
y'
w'
$C'_{S\text{-translated}}$
C'_T
•Identify potential matches between **S** and **T**
•Select the best translation for each class that matches the corresponding concept in **T**, where similarity score > θ.
•Output: matched terms (**m**) between **S** & **T**
x
y
z
w
a
c
d
f
$T_{enriched}$

Terminological Matching

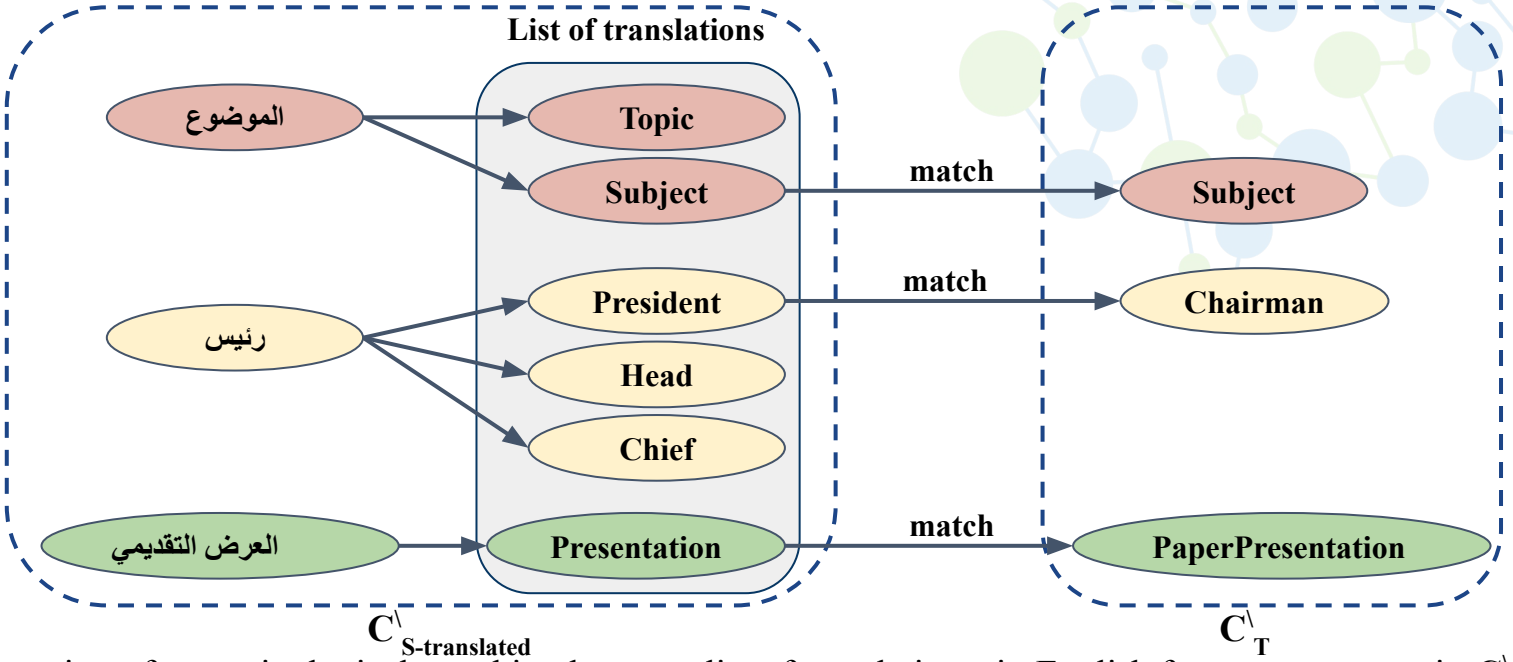


Fig. 2. Illustration of a terminological matching between list of translations, in English,for every concept in $C^{\setminus}_{\text{S-translated}}$, in Arabic, and $C^{\setminus}_{T}$ in English

Terminological Matching Algorithm

Algorithm 1: Terminological Matching

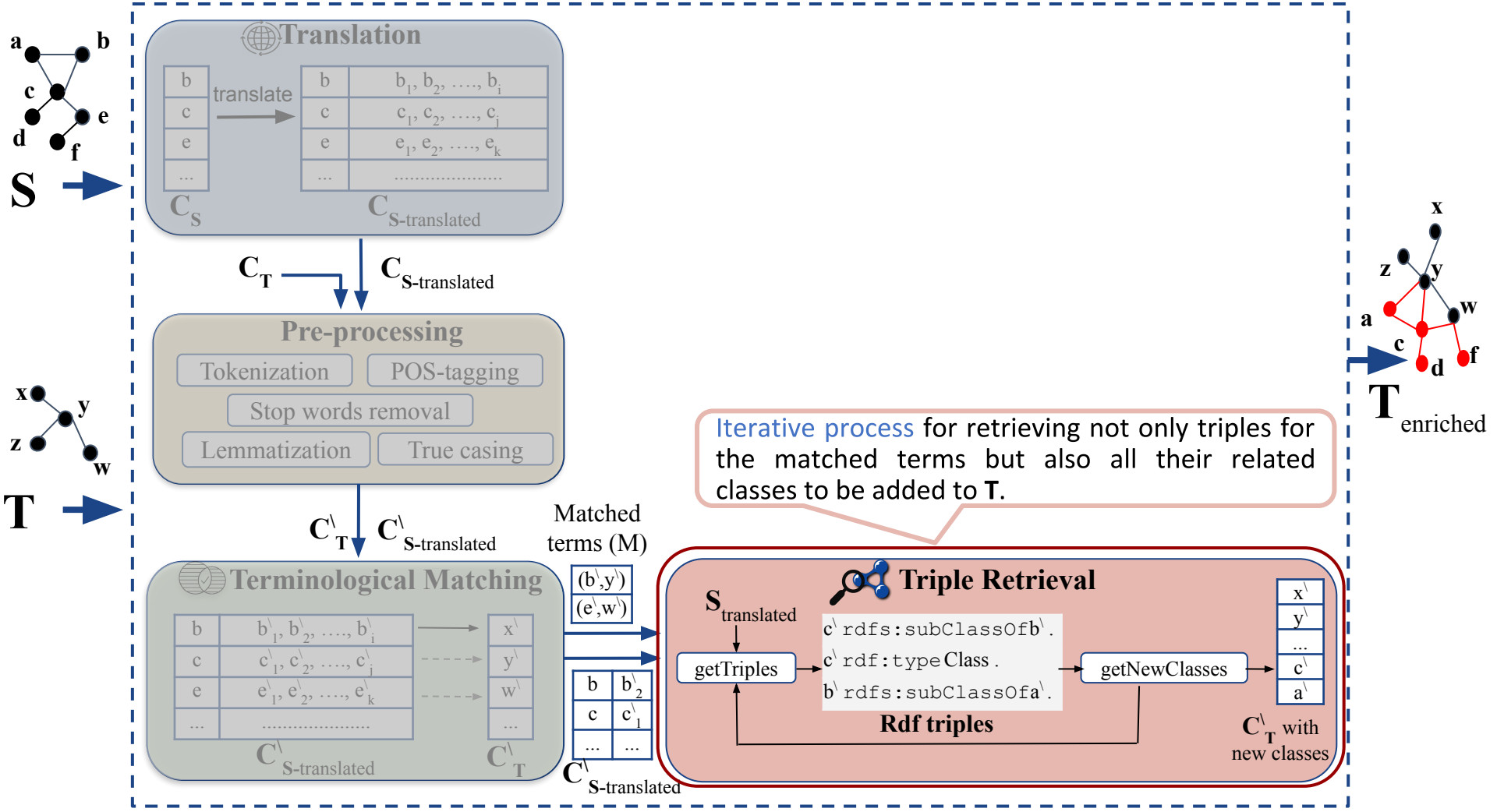
Data: $\mathcal{C}'_{S-translated}$, $\mathcal{C}'_T$, θ similarity threshold
Result: M matched terms, $\mathcal{C}'_{S-translated}$

```
foreach c_s ∈ C'_{S−translated}, t ∈ listOfTranslations, c_t ∈ C'_T do
    similarityScore ← getSimilarity(t,c_t)
    if similarityScore ≥ θ then
        M ::= (t, c_t)
        C'_{S−translated} = update(C'_{S−translated},M)
Function getSimilarity(sentence1, sentence2):double
    similarity ← getJaccardSimilarity(sentence1, sentence2)
    if similarity ≠ 1 then
        similarity ← (sentenceSimilarity(sentence1,sentence2)
        + sentenceSimilarity(sentence2,sentence1))/2
    return similarity
Function sentenceSimilarity(sentence1, sentence2):double
    simScore ← 0.0
    count ← 0.0
    foreach w_i ∈ sentence1.split(" ") do
        foreach w_j ∈ sentence2.split(" ") do
            pathSim ::= getPathSimilarity(w_i,w_j)
        simScore+ = pathSim.max
        count+ = 1
    simScore ← simScore/count
    return simScoure
```

Pairwise similarity between the list of translations of each class in $\mathbf{C}^{\setminus}_{\text{S-translated}}$ and $\mathbf{C}^{\setminus}_{\text{T}}$

Filter the identical concepts using Jaccard similarity

Calculate semantic similarity using the path length measure based on WordNet

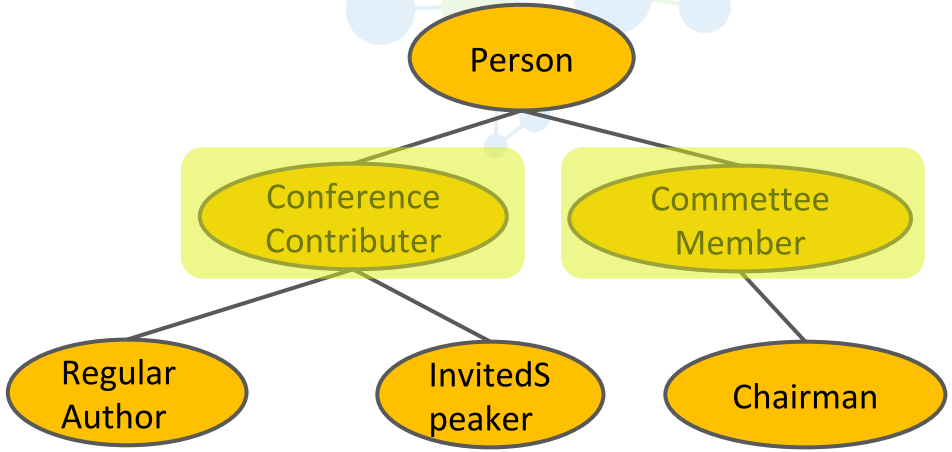

S
T
Translation
translate
C_S
$C_{S\text{-translated}}$
C_T
Pre-processing
Tokenization
POS-tagging
Stop words removal
Lemmatization
True casing
C'_T
$C'_{S\text{-translated}}$
Terminological Matching
$C'_{S\text{-translated}}$
C'_T
Matched terms (M)
(b',y')
(e',w')
b c e ...
$C'_{S\text{-translated}}$
Iterative process for retrieving not only triples for the matched terms but also all their related classes to be added to T.
Triple Retrieval
$S_{translated}$
getTriples
c' rdfs:subClassOf b' .
c' rdf:type Class .
b' rdfs:subClassOf a' .
Rdf triples
getNewClasses
x' y' ... c' a'
C'_T with new classes
$T_{enriched}$

Example

- For the matched term *m* = “person”, get all triples from $S_{translated}$ where *m* is subject/object.

1st iteration:

(conference contributor,subClassOf,person)
(committee member,subClassOf,person)



$S_{translated}$ = $Conference_{en}$

Example

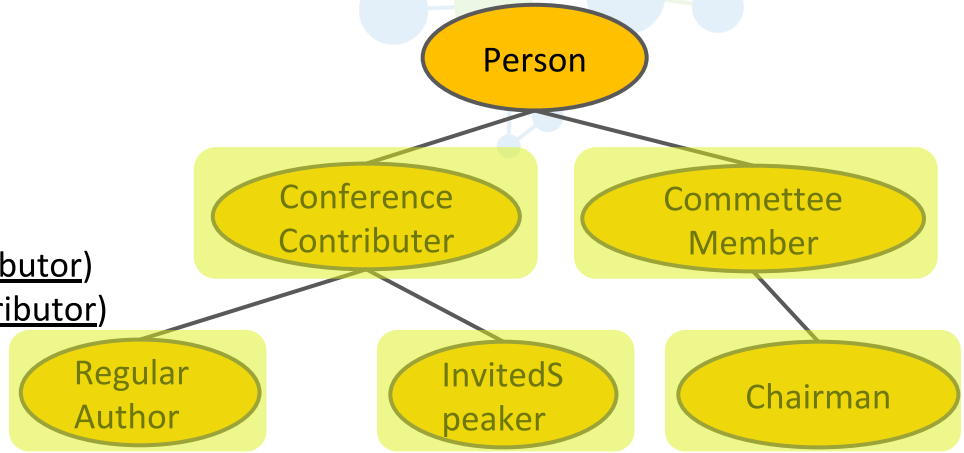
- For the matched term *m* = “person”, get all triples from $S_{translated}$ where *m* is subject/object.

1st iteration:

(conference contributor,subClassOf,person)
(committee member,subClassOf,person)

2nd Iteration:

(conference contributor,type,Class)
(regular author,subClassOf,conference contributor)
(invited speaker,subClassOf,conference contributor)
(committee member,type,Class)
(chairman,subClassOf,committee member)



$S_{translated}$ = $Conference_{en}$

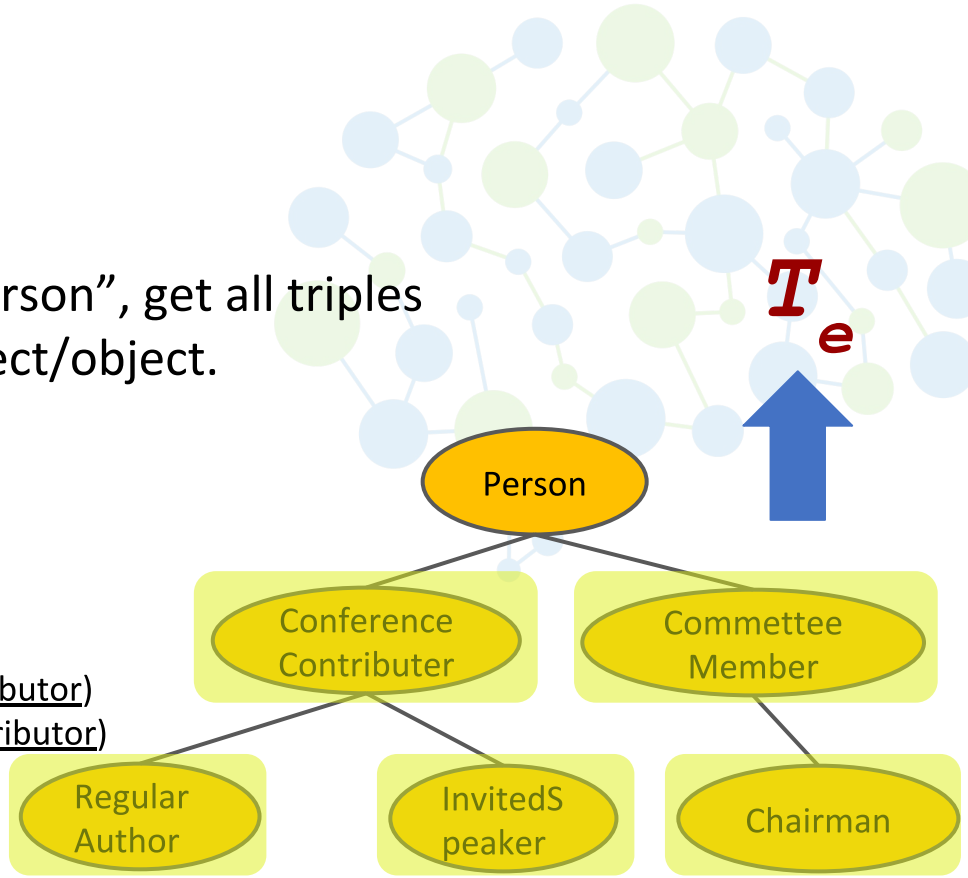
Example

- For the matched term *m* = “person”, get all triples from $S_{translated}$ where *m* is subject/object.

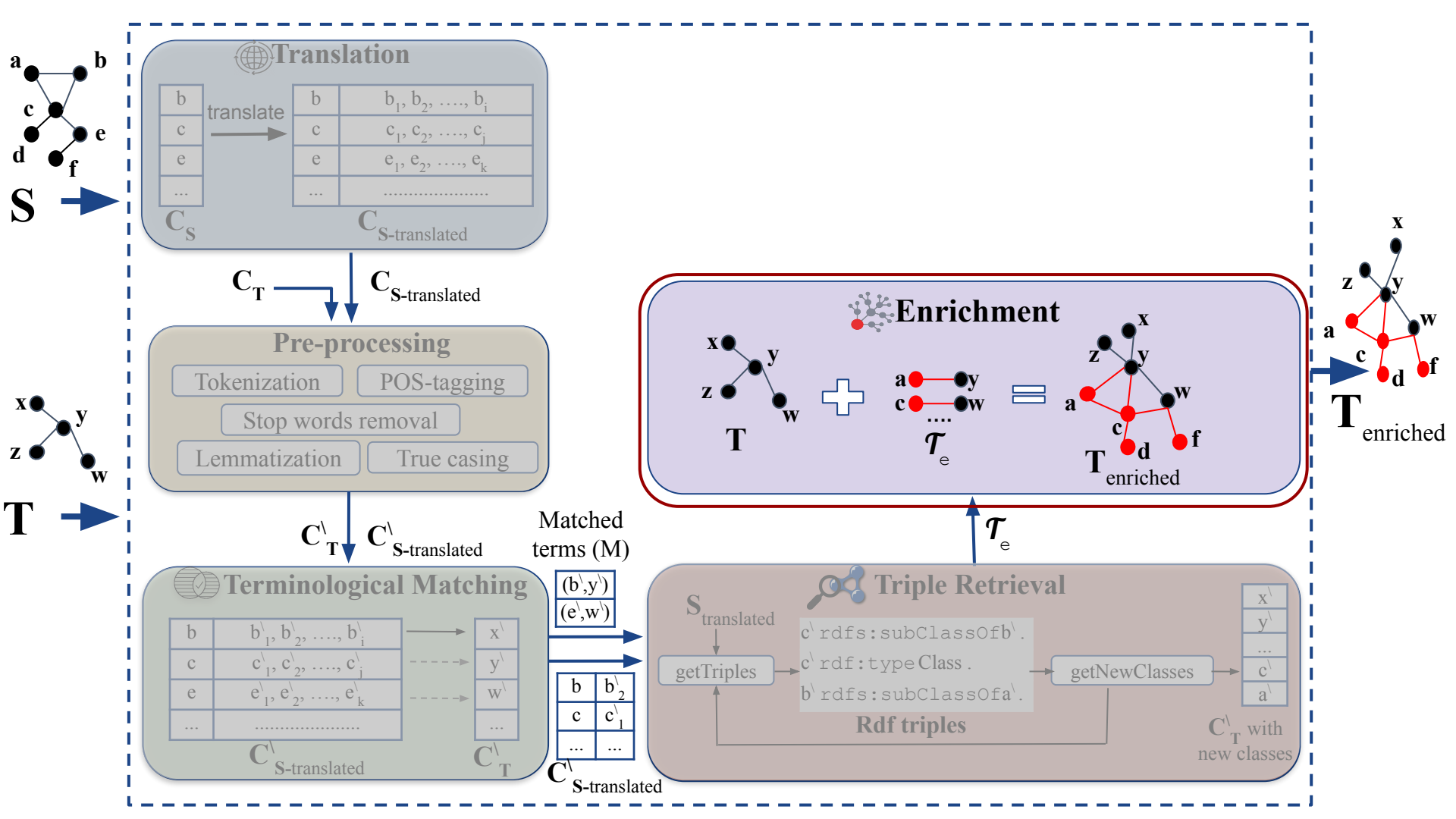
1st iteration:
(conference contributor,subClassOf,person)
(committee member,subClassOf,person)

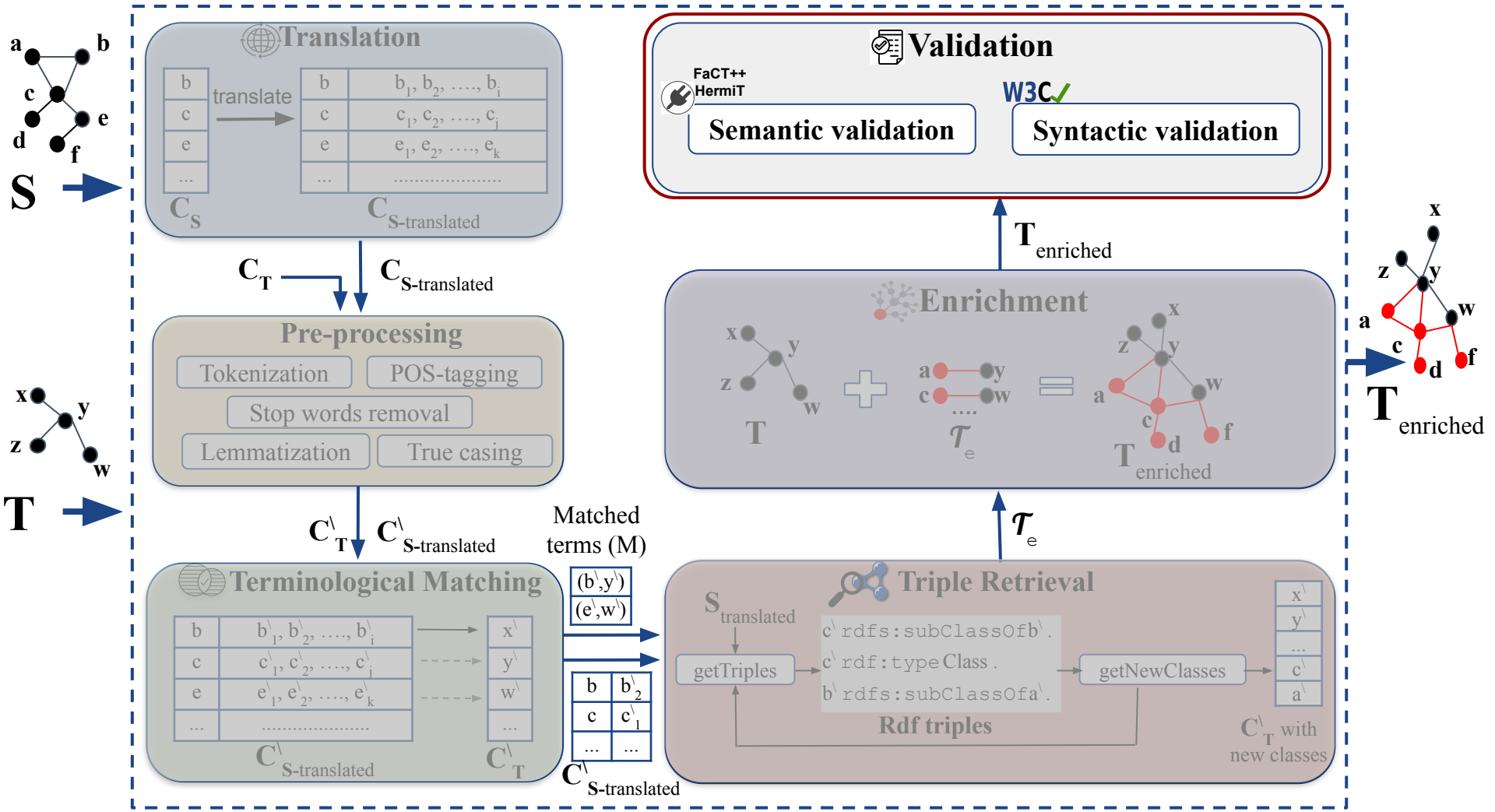
2nd Iteration:
(conference contributor,type,Class)
(regular author,subClassOf,conference contributor)
(invited speaker,subClassOf,conference contributor)
(committee member,type,Class)
(chairman,subClassOf,committee member)

nth Iteration:
None

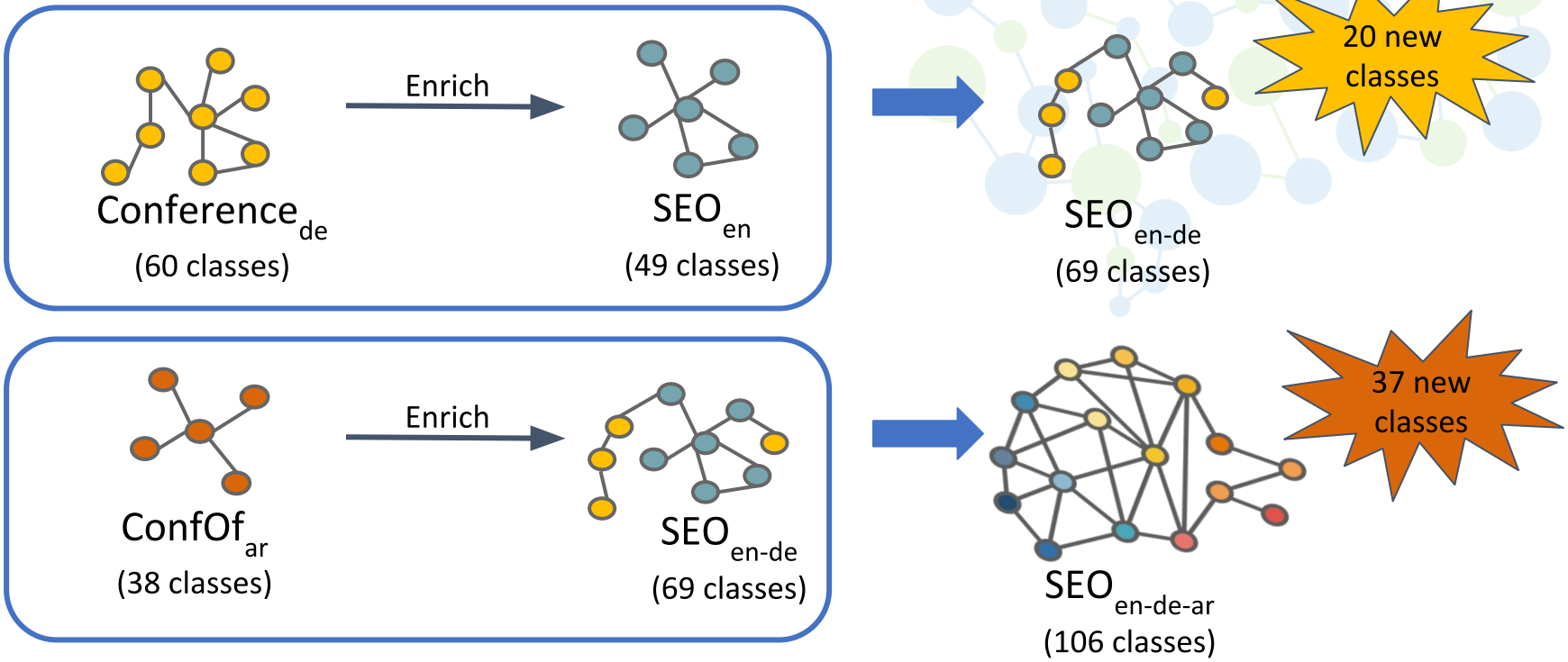


$S_{translated} = Conference_{en}$


S
T
Translation
translate
C_S
$C_{S\text{-translated}}$
C_T
Pre-processing
Tokenization
POS-tagging
Stop words removal
Lemmatization
True casing
C'_T
$C'_{S\text{-translated}}$
Terminological Matching
$C'_{S\text{-translated}}$
C'_T
Matched terms (M)
$C'_{S\text{-translated}}$
$S_{translated}$
getTriples
c' rdfs:subClassOf b' .
c' rdf:type Class .
b' rdfs:subClassOf a' .
Rdf triples
getNewClasses
C'_T with new classes
Triple Retrieval
τ_e
Enrichment
T
τ_e
$T_{enriched}$
$T_{enriched}$


S
T
Translation
translate
C_S
$C_{S\text{-translated}}$
C_T
Pre-processing
Tokenization
POS-tagging
Stop words removal
Lemmatization
True casing
C'_T
$C'_{S\text{-translated}}$
Terminological Matching
$C'_{S\text{-translated}}$
C'_T
Matched terms (M)
(b',y')
(e',w')
b c e ...
$C'_{S\text{-translated}}$
$S_{translated}$
getTriples
c' rdfs:subClassOf b'.
c' rdf:type Class .
b' rdfs:subClassOf a'.
Rdf triples
getNewClasses
C'_T with new classes
Triple Retrieval
τ_e
Enrichment
T
τ_e
$T_{enriched}$
Validation
FaCT++
HermiT
Semantic validation
W3C
Syntactic validation
$T_{enriched}$

Use case: Enriching the Scientific Events Ontology (SEO)



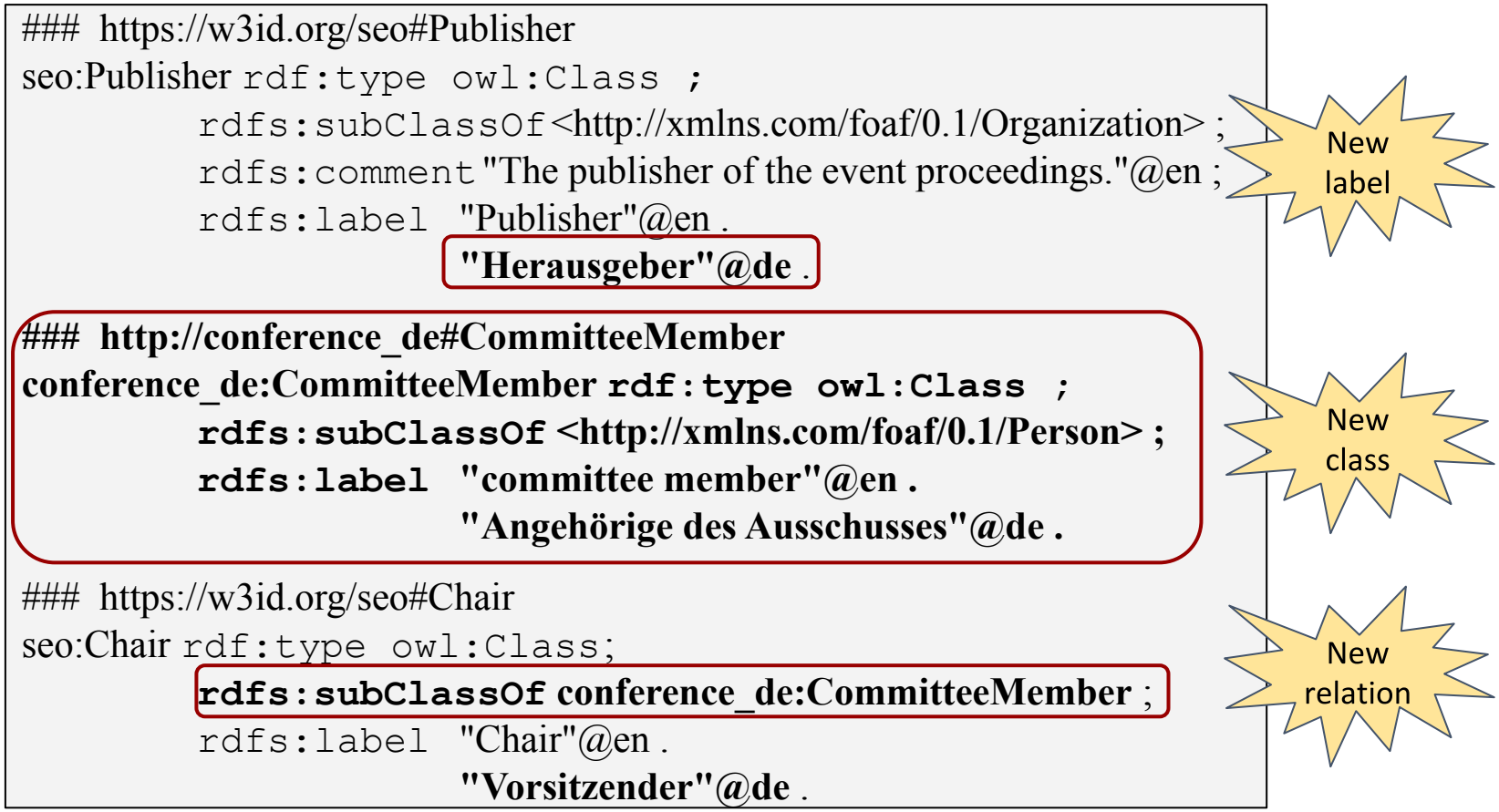


Fig. 3. Small fragment from SEO_{en-de} ontology after the enrichment.

Evaluation

Dataset:

- MultiFarm benchmark
  - Designed for evaluating cross-lingual ontology matching systems
  - Consists of:
    - Seven ontologies originally coming from the Conference benchmark
    - Their translation into nine languages
    - The corresponding cross-lingual alignments between them.

Evaluation

Dataset:

- MultiFarm benchmark
  - Designed for evaluating cros systems
  - Consists of:
    - Seven ontologies originally benchmark
    - Their translation into nine languages
    - The corresponding cross-lingual alignments between them.

Ontology Name	Classes	Datatype Properties	Object Properties
Conference	60	18	46
Sigkdd	49	11	17
Iasted	140	3	38
ConOf	38	23	13
Cmt	36	10	49
Ekaw	74	0	33
Edas	104	20	30

Evaluation


Effectiveness of OECM
Tasks
Enrichment Process Quality
Comparison with state-of-the-art

Evaluation

Effectiveness of OECM

- Comparing results with the reference alignment.

Table 2. Precision, recall and F-measures for the cross-lingual matching

Ontology pairs	German× English			Arabic × English					
	Precision	Recall	F-measure	Precision		Recall		F-measure	
				Before	After	Before	After	Before	After
Conference × Cmt	1.00	0.38	0.56	1.00	1.00	0.33	0.42	0.50	0.59
ConfOf × Cmt	1.00	0.70	0.82	1.00	1.00	0.30	0.60	0.46	0.75
Sigkdd × Cmt	1.00	0.90	0.95	1.00	1.00	0.40	0.80	0.57	0.89

After linguistic correction

Evaluation

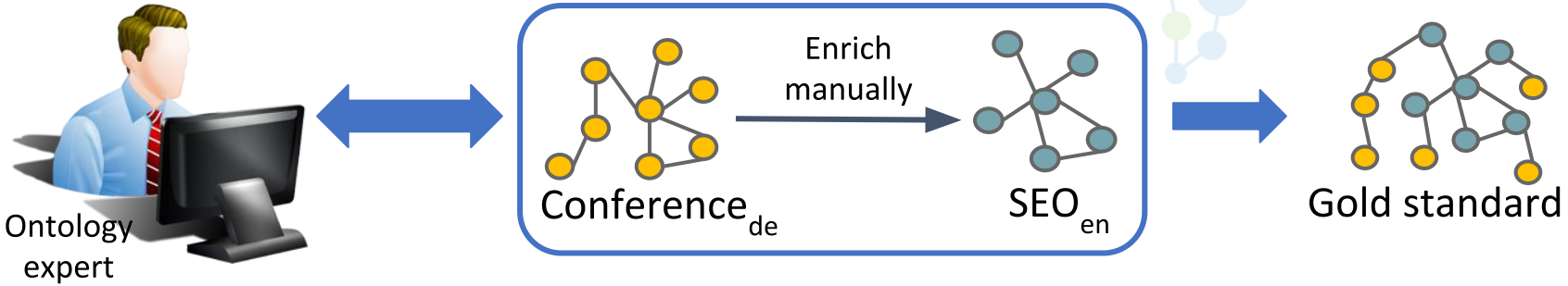
Comparison with the state-of-the-art

Table 3. State-of-the-art comparison results.

Approaches	$Conference_{de}$ × $Ekaw_{en}$			$Conference_{de}$ × $Edas_{en}$		
	Precision	Recall	F-measure	Precision	Recall	F-measure
AML [7]	0.56	0.20	0.29	0.86	0.35	0.50
KEPLER [16]	0.33	0.16	0.22	0.43	0.18	0.25
LogMap [15]	0.71	0.20	0.31	0.71	0.29	0.42
XMap [28]	0.18	0.16	0.17	0.23	0.18	0.20
OECM 1.0 [14]	0.75	0.67	0.71	0.93	0.76	0.84
OECM 1.1	**1.00**	**0.80**	**0.89**	**1.00**	**0.78**	**0.88**
	$Conference_{ar}$ × $Ekaw_{en}$			$Conference_{ar}$ × $Edas_{en}$		
AML [7]	0.64	0.39	0.28	0.71	0.42	0.29
KEPLER [16]	0.40	0.30	0.24	0.40	0.30	0.24
LogMap [15]	0.40	0.13	0.08	0.40	0.18	0.12
XMap [28]	**1.00**	0.0	0.0	**1.00**	0.00	0.00
OECM 1.1	**1.00**	**0.50**	**0.67**	0.86	**0.67**	**0.75**

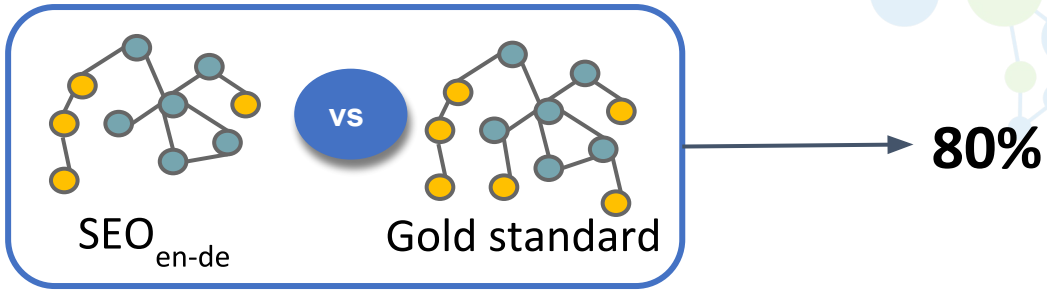
Evaluation

Evaluating the Enrichment Process Quality



Evaluation

Evaluating the Enrichment Process Quality



Conclusion

- OECM (fully automated approach) creates multilingual ontologies from monolingual ones.
- Indo and non-Indo-European languages resources are used for enrichment.
- Considering multiple translations of concepts and the use of semantic similarity measures have significantly improved the quality of the matching process.

Conclusion

- Iterative triple retrieval process has been developed to retrieve not only triples for the matched terms but also all their related classes, to be added to ***T***.
- The results of the cross-lingual matching process are found promising compared to five state-of-the-art approaches.
- The linguistic corrections for the Arabic ontologies considerably enhanced the matching results.

Future Work

- Consider properties and individuals in the enrichment process.
- Apply optimization methods in order to evaluate the efficiency of OECM when enriching very large ontologies.

From Monolingual to Multilingual Ontologies: The Role of Cross-lingual Ontology Enrichment

Thank You!

Shimaa Ibrahim
ibrahim@cs.uni-bonn.de