AN ONTOLOGY ALIGNMENT APPROACH COMBINING WORD EMBEDDING AND THE RADIUS MEASURE

Molka Tounsi Dhouib
Catherine Faron Zucker
Andrea G. B. Tettamanzi

Université Côte d'Azur, Inria, CNRS, I3S,
Sophia Antipolis, France
Silex, France

SEMANTiCS
Karlsruhe 2019

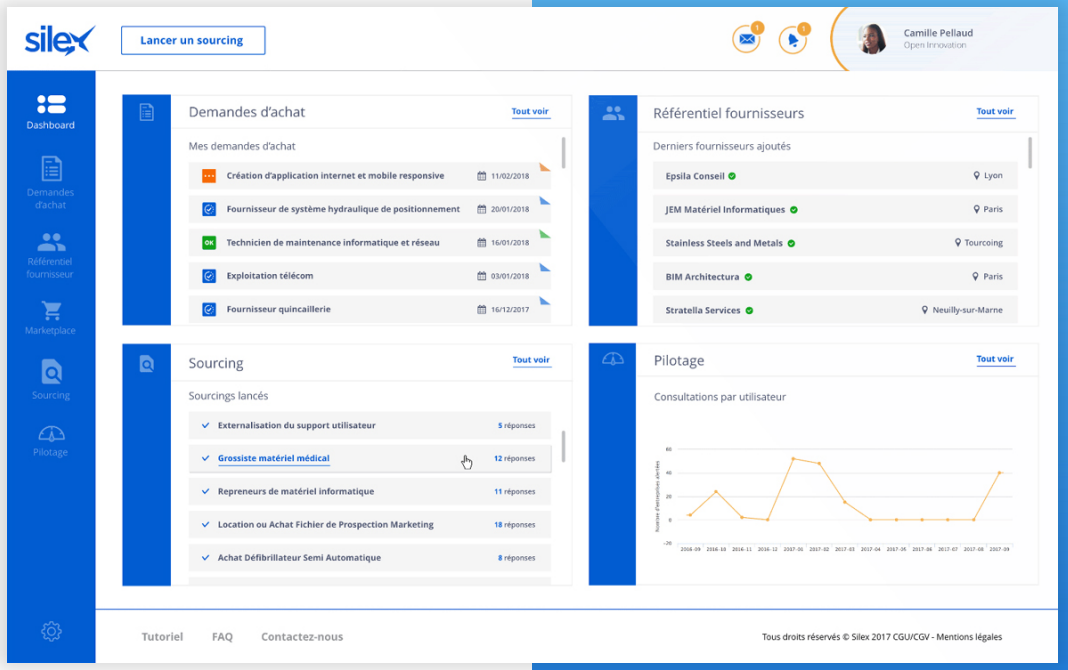
www.silex-france.com

THE COMPANY

Silex offers tomorrow's **SaaS solutions** for optimized sourcing.

PRODUCT

B2B platform to automate supplier identification and recommend companies for purchasing projects.



TECHNOLOGY

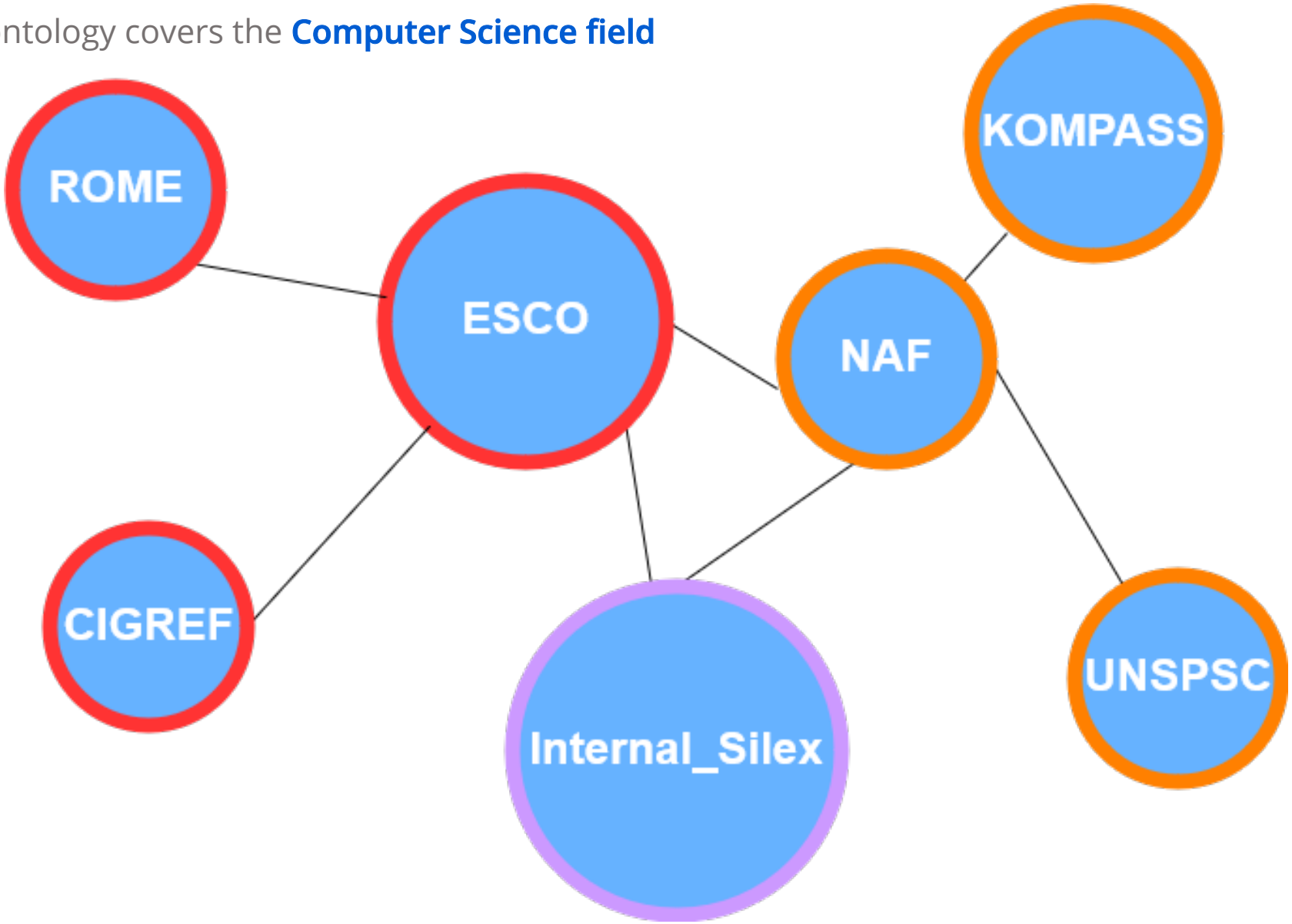
Artificial Intelligence
Semantic Web

CLIENTS

Our customers are large companies in both the public and private sectors.

MOTIVATION

- Currently Silex ontology covers the **Computer Science field**



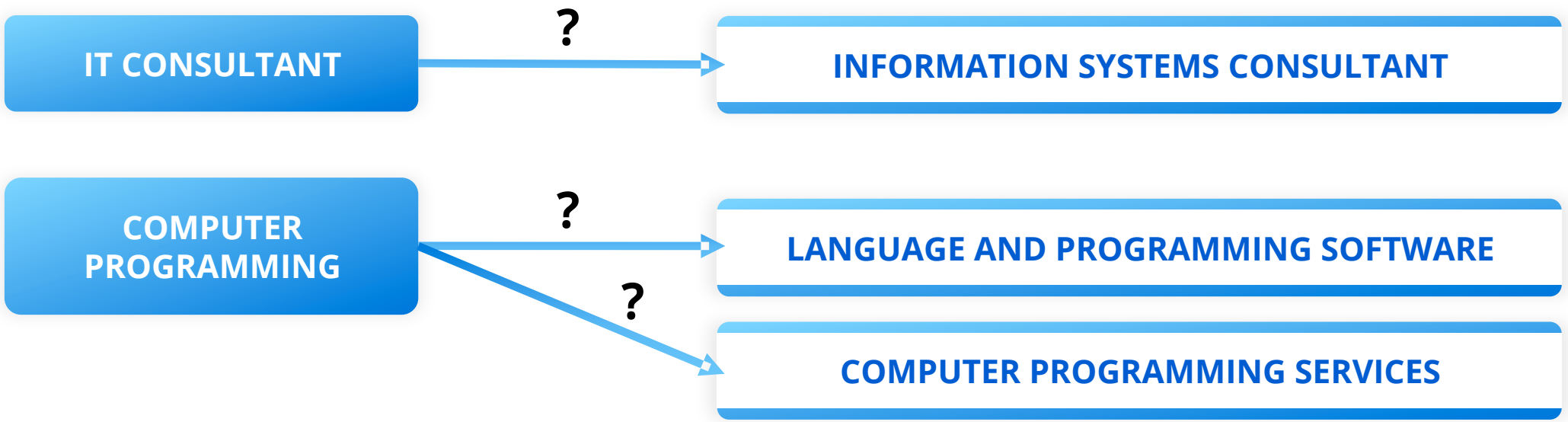
HOW ? CHALLENGES ?

HOW

How to automatically align the entire vocabularies to extend the Silex ontology to all business sectors?

CHALLENGES

- Dealing with heterogeneity
- Reuse information between ontologies



WHAT DO WE NEED ?

- Define a measure of the similarity between entities or concepts of two ontologies
- Define a methodology to refine the nature of the relationship between two similar concepts
  - Equivalence relation: **skos:closeMatch/owl:sameAs**
  - Hierarchical relation: **skos:broadMatch, skos:narrowMatch / rdfs:subClassOf, rdfs:subPropertyOf**

RELATED WORKS ON ONTOLOGY ALIGNMENT

ELEMENT-LEVEL TECHNIQUES

Calculating the surface similarity between lexical information of entities (labels, comments ...)

STRUCTURE-LEVEL TECHNIQUES

Analysis of the neighbourhood of two entities

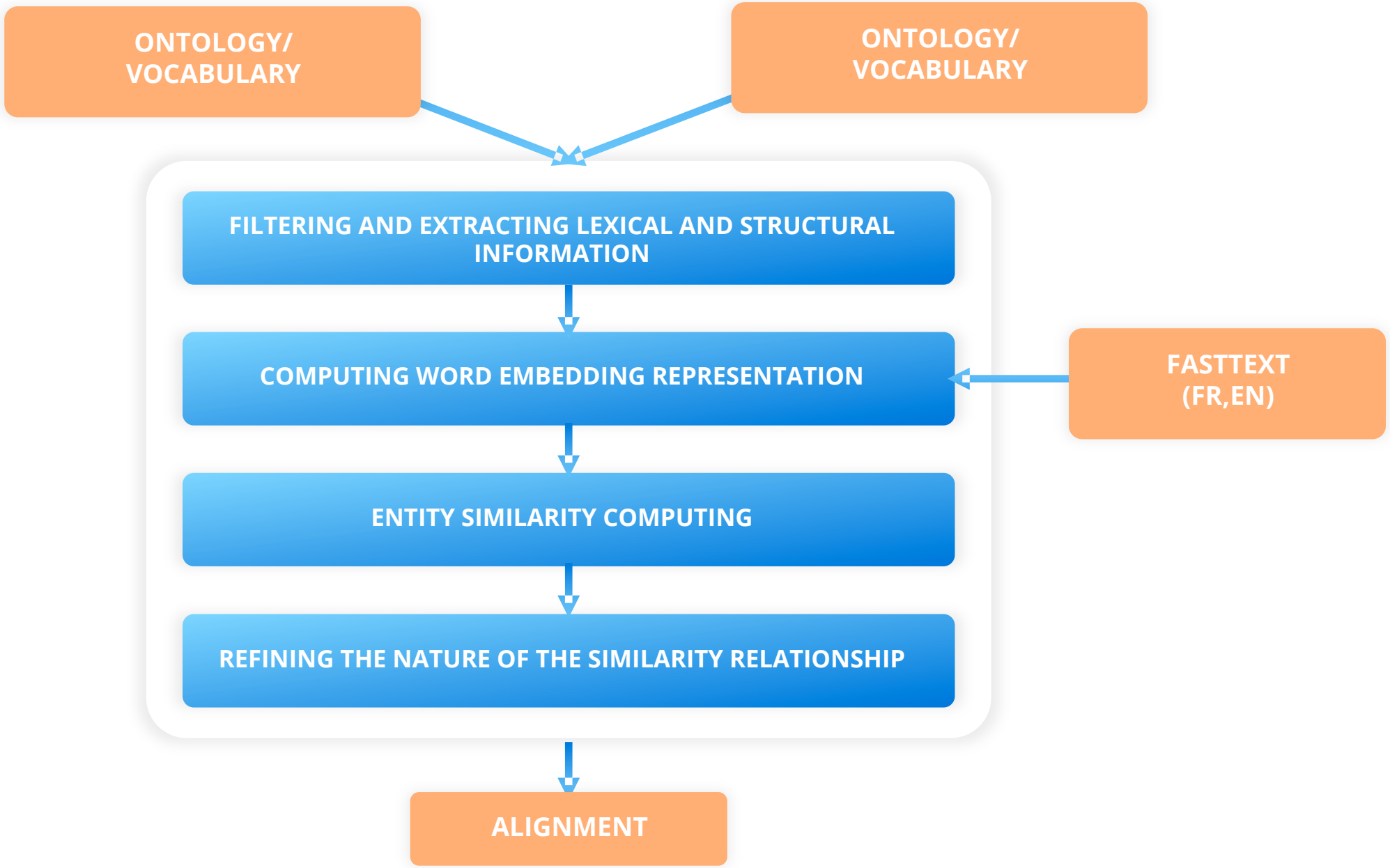
EXTERNAL TECHNIQUES

Use external information sources of a domain (wordnet, Wikipedia)

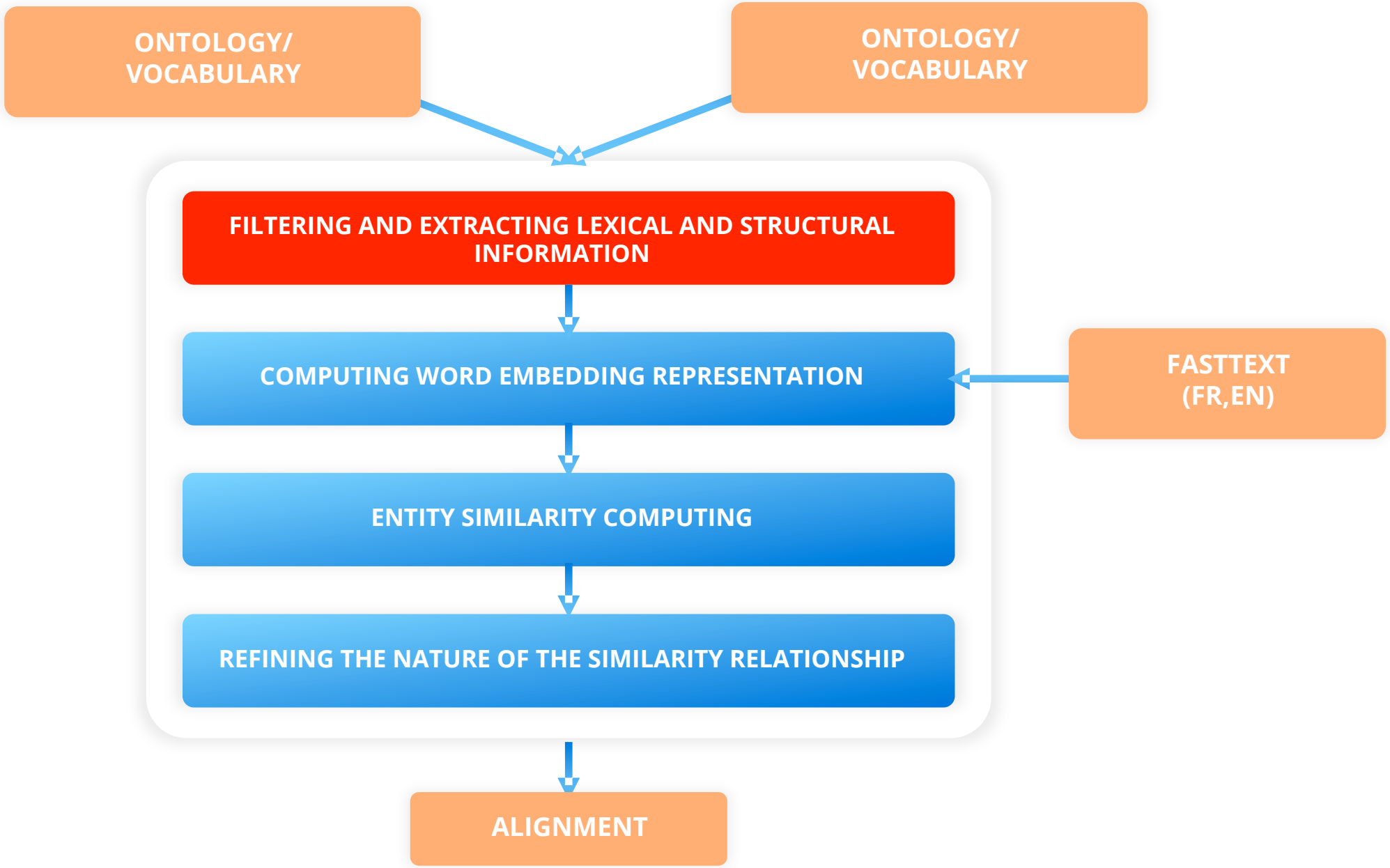
SEMANTIC TECHNIQUES

Interpret the meaning of the entities (word embeddings)

OVERVIEW OF OUR APPROACH OF ONTOLOGY ALIGNMENT



OVERVIEW OF OUR APPROACH OF ONTOLOGY ALIGNMENT



OVERVIEW OF OUR APPROACH TO ONTOLOGY ALIGNMENT

EXTRACTING LEXICAL AND STRUCTURAL INFORMATION

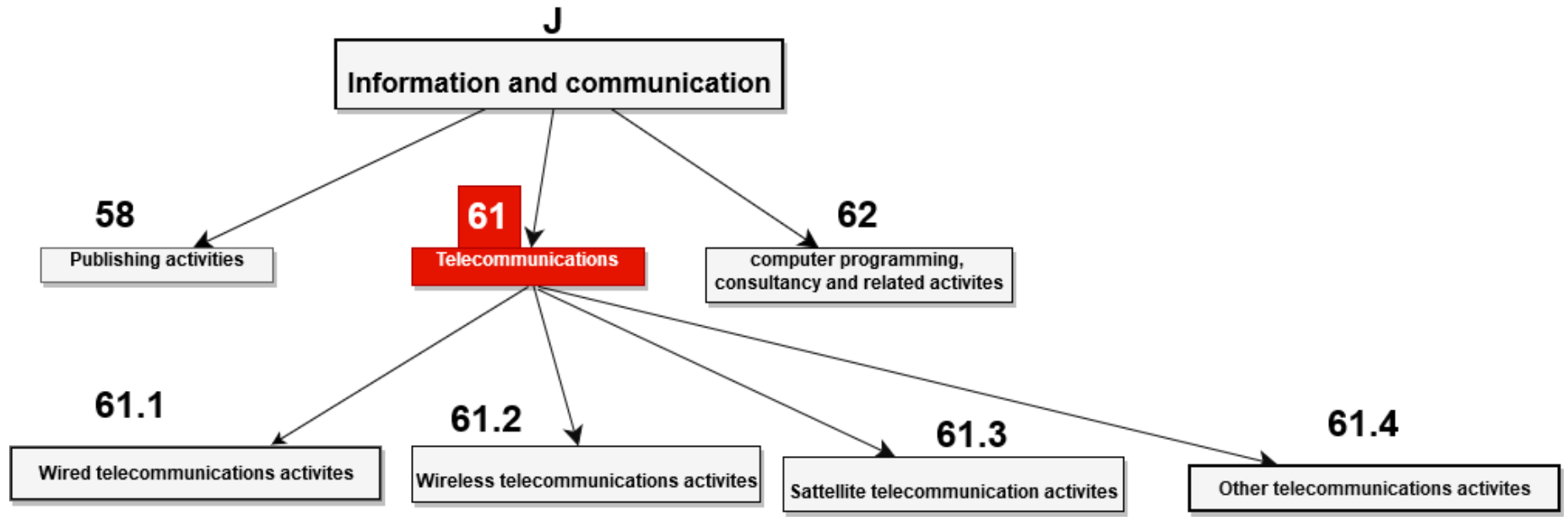
LEXICAL INFORMATION

STRUCTURAL INFORMATION

```
SELECT ?uri ?label
 (group_concat(DISTINCT ?mid_label; separator=":")  AS ?lineage)
WHERE {
  ?uri skos:prefLabel/rdfs:label ?label
  FILTER (lang(?label)='fr’ en)
  ?uri ^skos:broader* rdfs:subClass/rdfs:subproperties  ?mid .
  ?mid skos:prefLabel/rdfs:label ?mid_label .
  FILTER (lang(?mid_label)=‘fr’ en)
} GROUP BY ?mid ORDER BY count(?label)
```

OVERVIEW OF OUR APPROACH TO ONTOLOGY ALIGNMENT

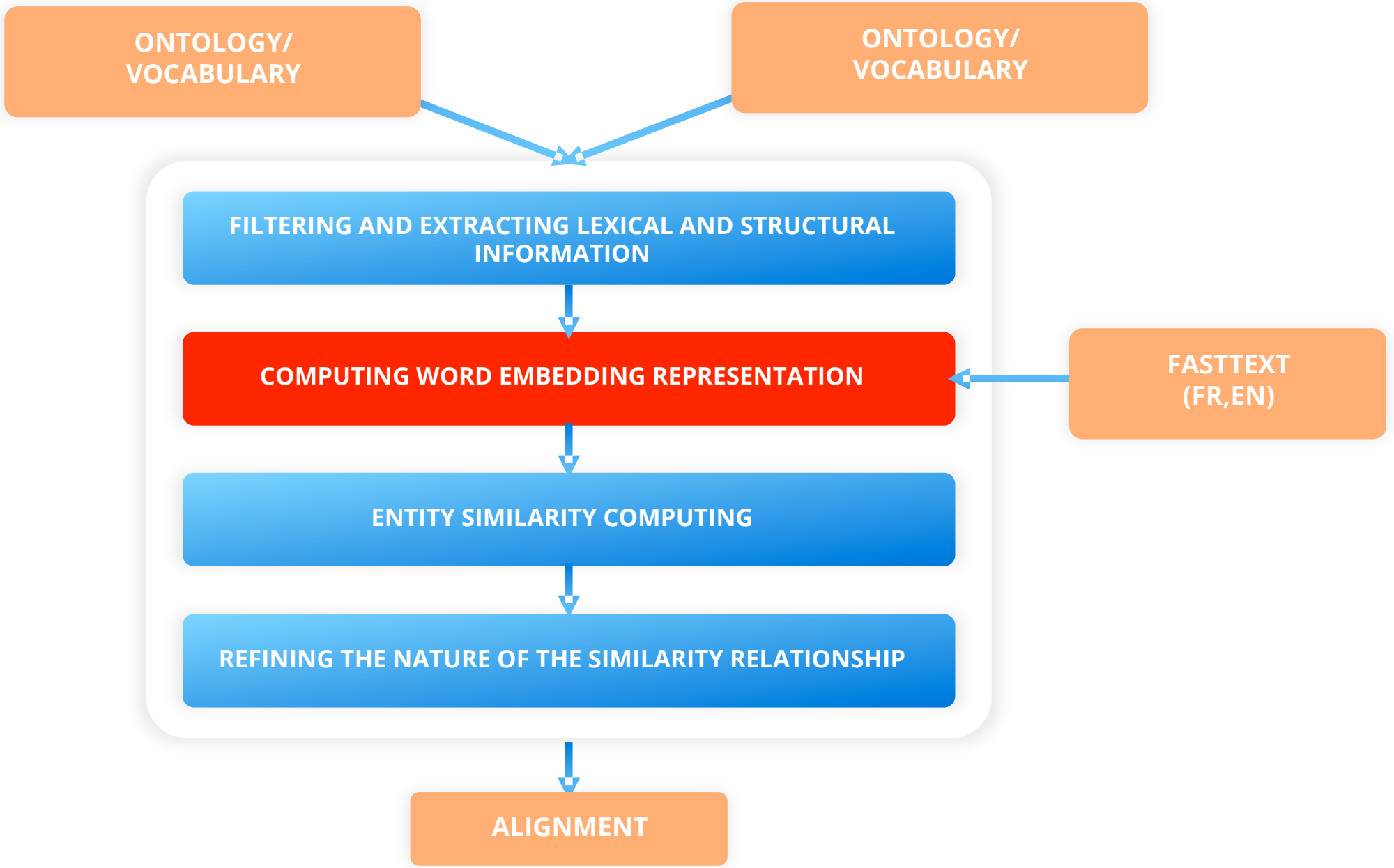
EXTRACTING LEXICAL AND STRUCTURAL INFORMATION



Lexical information (#61) ={Telecommunications}

Structural information (#61) ={Telecommunications, Wired telecommunications activities, Wireless telecommunications activities, Satellite telecommunication activities, Other telecommunications activities})

OVERVIEW OF OUR APPROACH OF ONTOLOGY ALIGNMENT

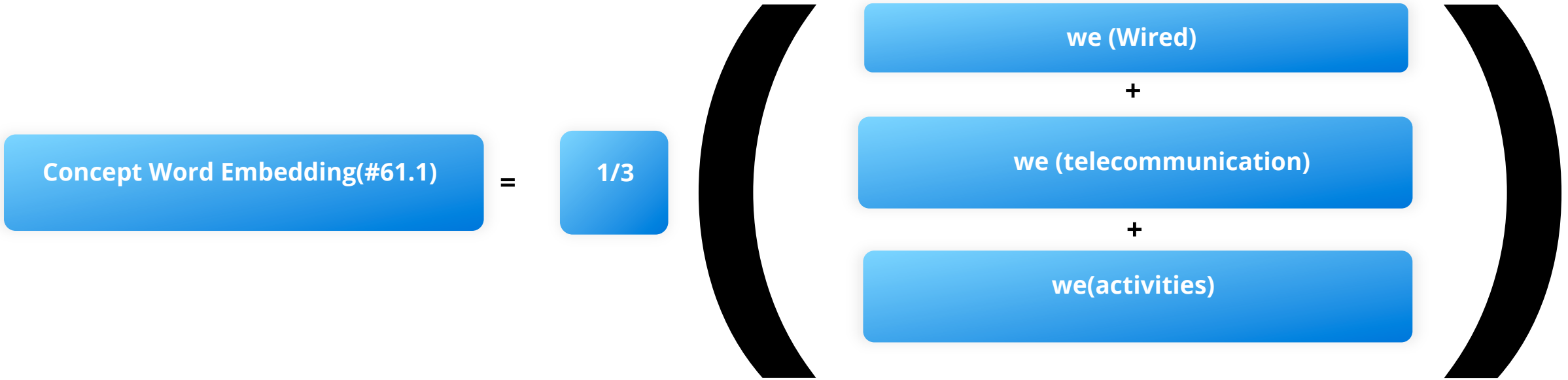


OVERVIEW OF OUR APPROACH TO ONTOLOGY ALIGNMENT

COMPUTING WORD EMBEDDING REPRESENTATIONS

Vector representation of a concept (lexical information)

$$\text{conceptWordEmbedding}(c) = 1/n \sum_{i=1}^{n} wi$$

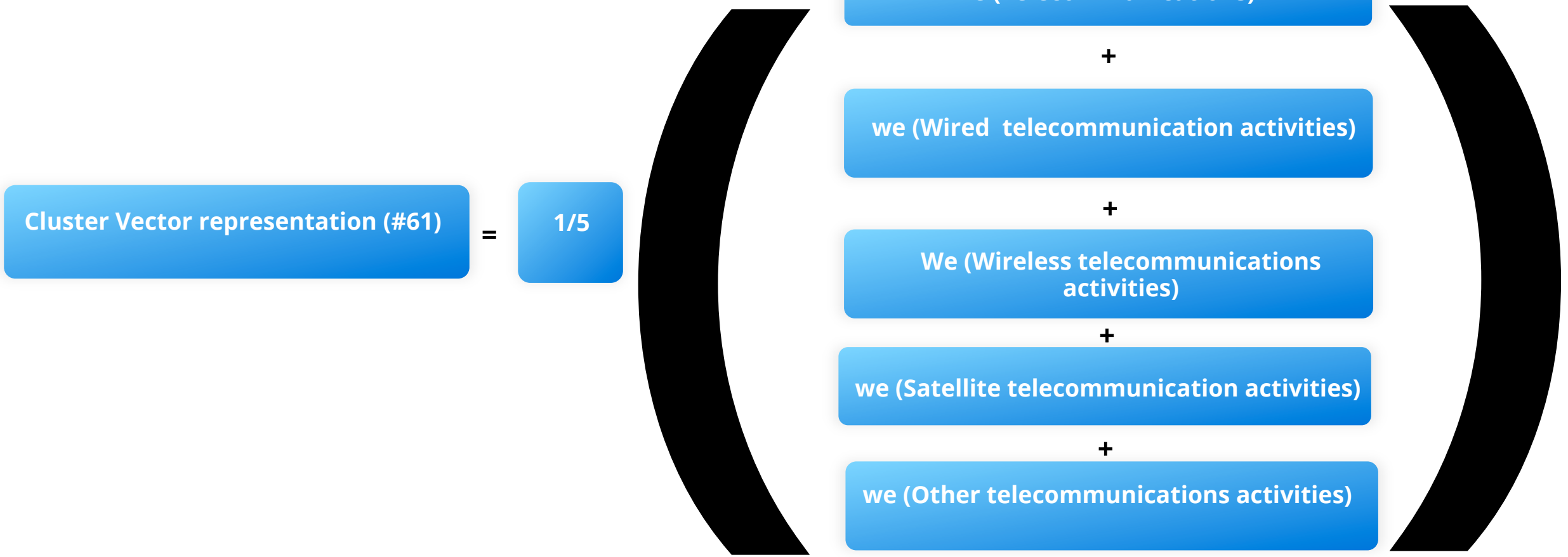


OVERVIEW OF OUR APPROACH TO ONTOLOGY ALIGNMENT

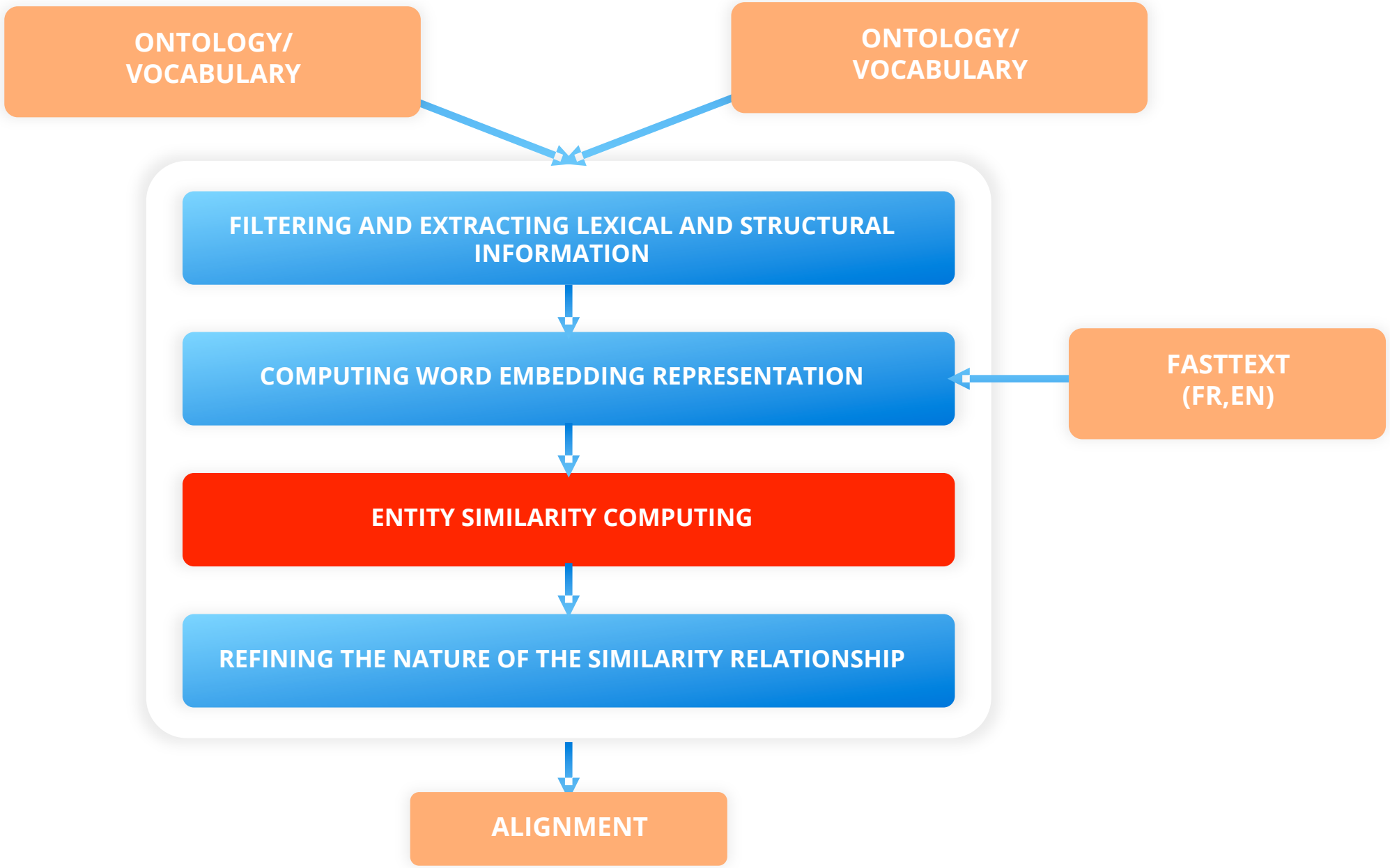
COMPUTING WORD EMBEDDING REPRESENTATIONS

Vector representation of a cluster (structural information)

$$clusterWordEmbedding(cl) = 1/k \sum_{i=1}^{k} conceptWordEmbedding(ci)$$

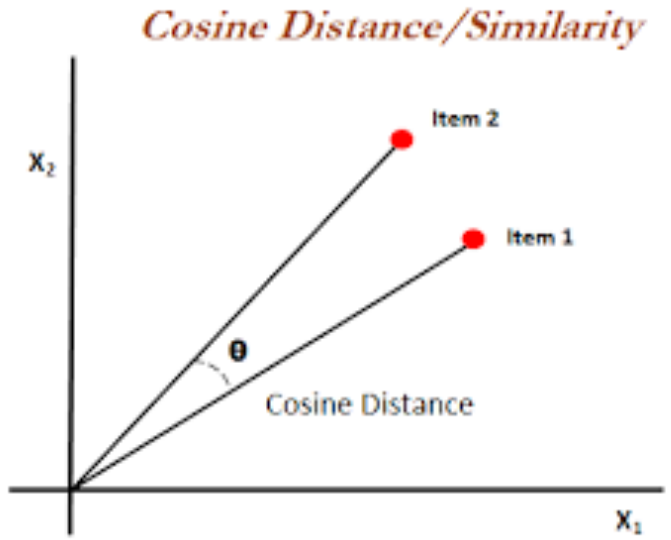


OVERVIEW OF OUR APPROACH OF ONTOLOGY ALIGNMENT

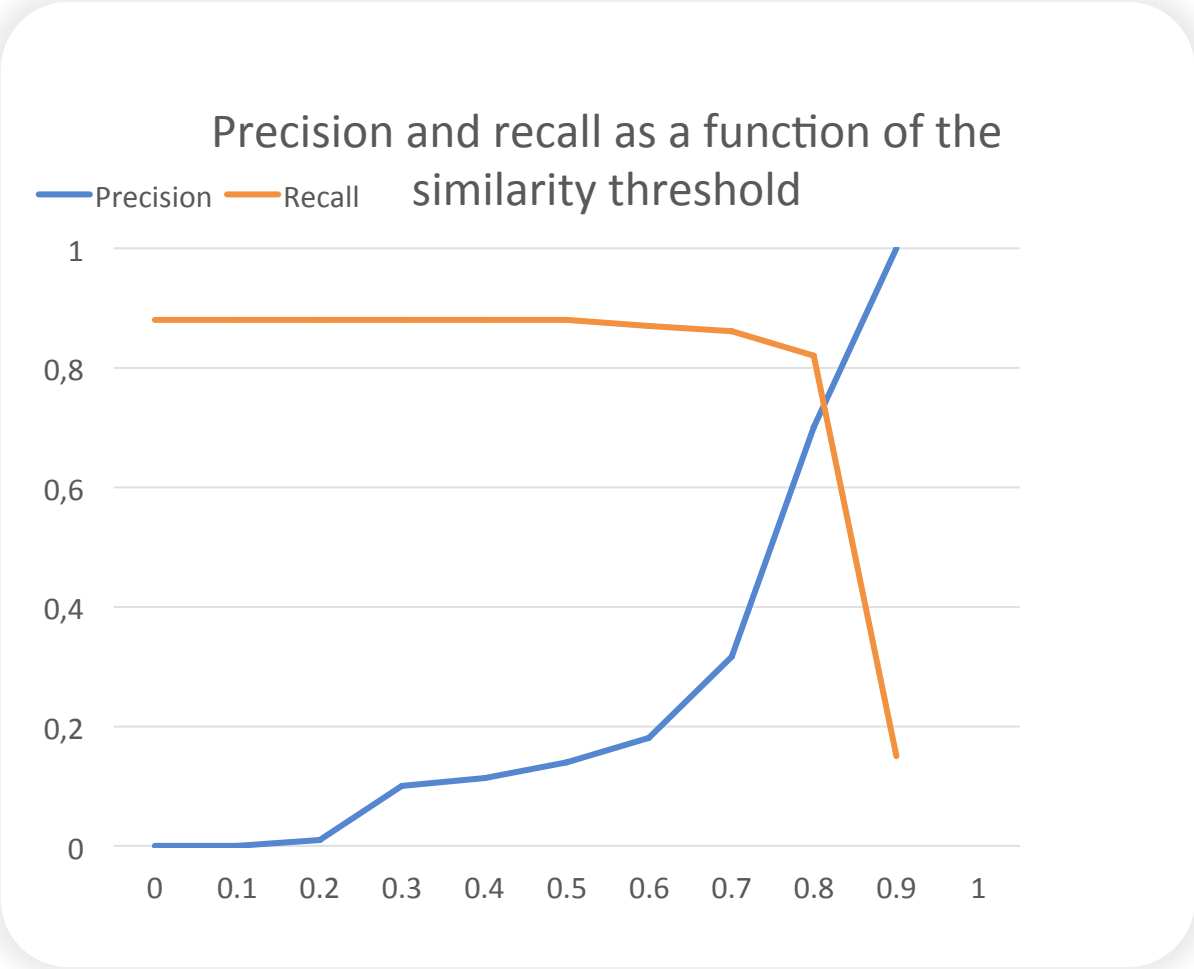


OVERVIEW OF OUR APPROACH TO ONTOLOGY ALIGNMENT

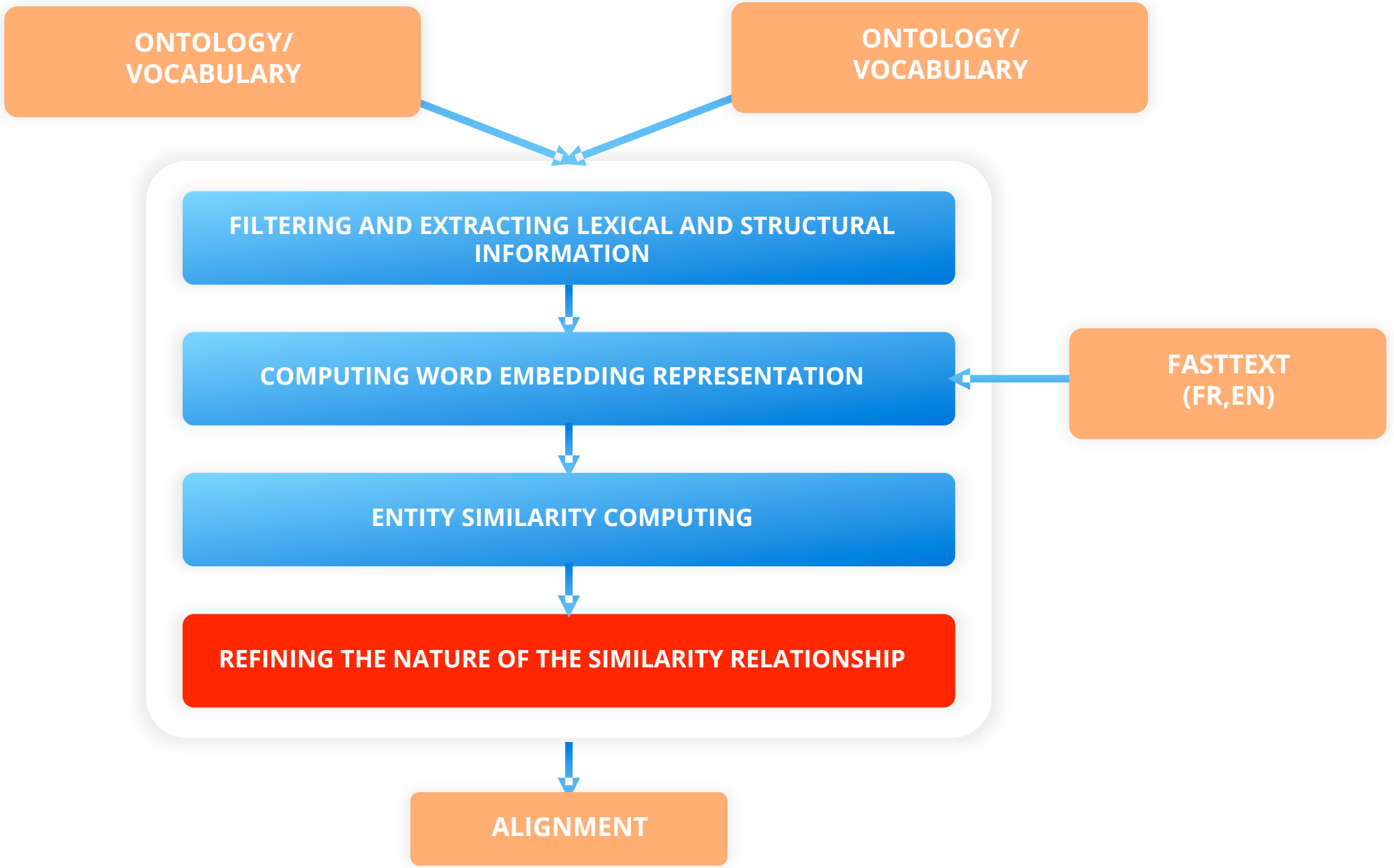
SEARCHING FOR MATCHING CONCEPTS



match (c1, c2) = cosine distance (c1, c2) > threshold

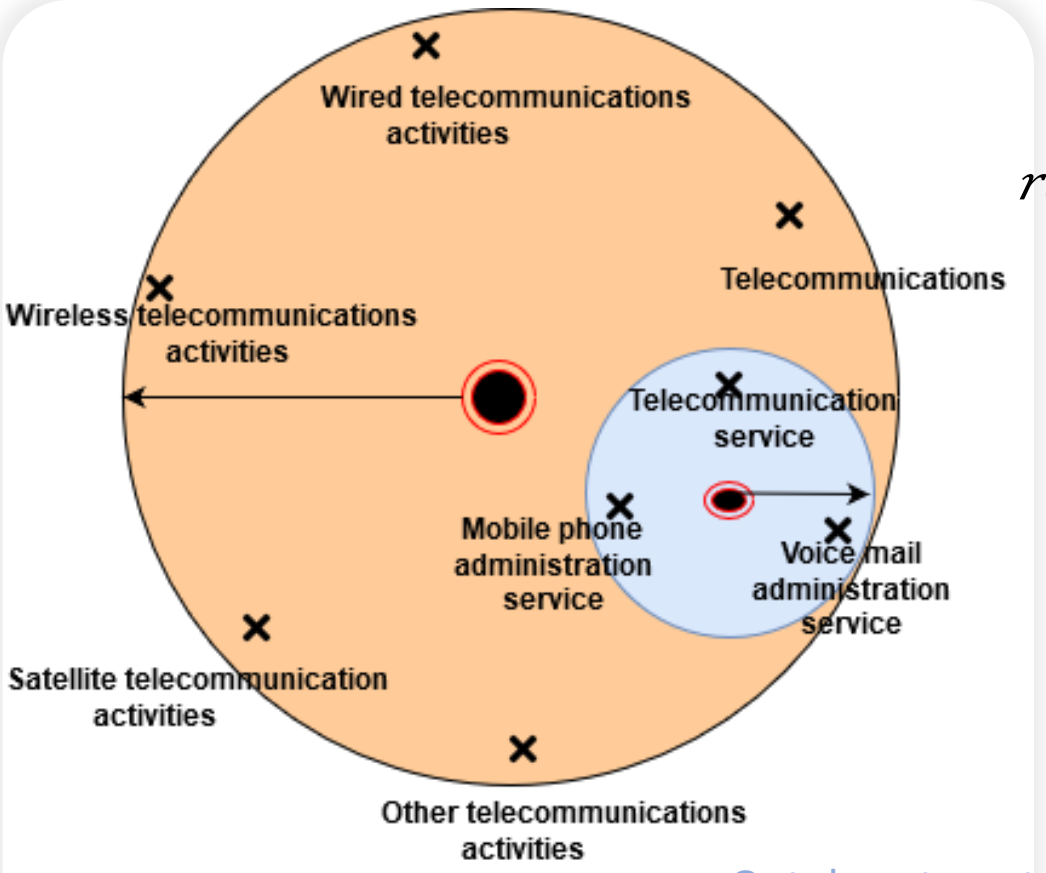


OVERVIEW OF OUR APPROACH OF ONTOLOGY ALIGNMENT



OVERVIEW OF OUR APPROACH TO ONTOLOGY ALIGNMENT

REFINING THE NATURE OF THE RELATIONSHIP BETWEEN TWO MATCHING CONCEPTS



$$radius = \sqrt{1/N \sum_{i=1}^{N} (1 - w_i \cdot w / |w_i| \cdot |w|)^2}$$

$$|\text{radius}(C_1) - radius(C_2)| < 0.1 \Rightarrow C_1 \text{ closeMatch } C_2$$

$$|\text{radius}(C_1) - radius(C_2)| > 0.1 \Rightarrow C_1 \text{ narrowMatch } C_2 \text{ and } C_2 \text{ broadMatch } C_1$$

EXPERIMENTS AND DISCUSSION

DATA

Task-Oriented Complex Alignment on Conference Organisation

- OAEI
- OWL format
- 57 complexe alignement

Silex use case

- Computer science field
- Gold standard provided by an expert in the Silex company

	classes	Object Properties	Data Properties
cmt	**30**	**49**	**10**
conference	**60**	**46**	**18**
confOf	**39**	**13**	**23**
edas	**104**	**30**	**20**
ekaw	**47**	**33**	**0**

Skills and Occupations		Business sectors	
ESCO	160	NAF	53
ROME	117	Kompass	574
Cigref	42	Silex	14

EXPERIMENTS AND DISCUSSION

EVALUATION PROTOCOL

Task-Oriented Complex Alignment on Conference Organisation

- Only complex alignement
- If the correct match is found among a proposed list, we consider that the entire proposed list is correct
- Precision, Recall and F-measure

Silex use case

- Precision, Recall and F-measure

EXPERIMENTS AND DISCUSSION

EXPERIMENTS ON TASK-ORIENTED COMPLEX ALIGNMENT ON CONFERENCE ORGANISATION

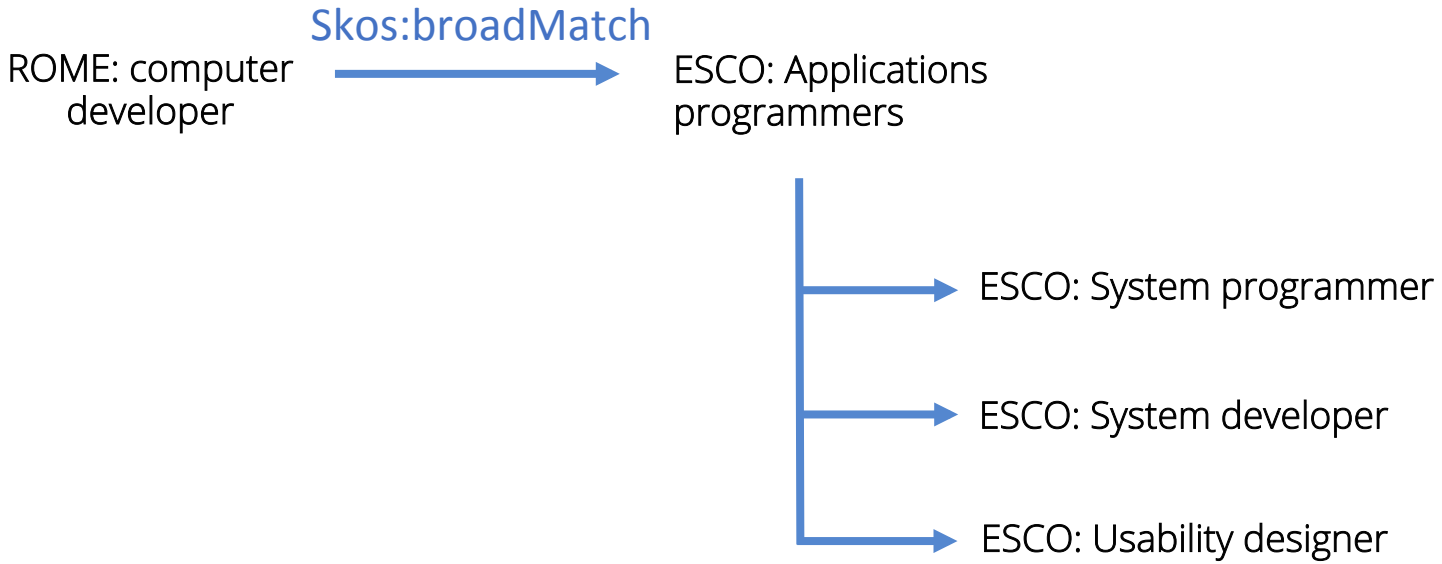
Systems	Precision	Recall	F-measure	Entities
Our System	**0.89**	**0.69**	**0.77**	All entities
Ritz e2009	0.30	0.13	0.19	All entities
Ritz e2010	0.83	0.09	0.18	All entities
Jiang 2016	0.09	0.11	0.10	All entities

- Cosine similarity < threshold ： cosine similarity ('chair main', 'demo chair') = 0.3 <0.8
- Our system is not designed to test hierarchical relations between two leaf nodes
- Assign equivalence relation instead of hierarchical relation

EXPERIMENTS AND DISCUSSION

EXPERIMENTS ON THE SILEX USE CASE (COMPUTER SCIENCE FIELD)

Relations	Precision	Recall
closeMatch	0.71- 0.80	0.60 – 0.95
narrowMatch	0.71- 0.83	0.69 – 1.00
broadMatch	0.73 -1.00	0.68 – 1.00



FUTURE WORK

WORD EMBEDDING

- Defining a specific set of pre-trained word vectors that best covers the Silex B2B use case
- Using the multilingual model of word embeding from fastText

RADIUS

- Performing an empirical study to define the optimal threshold for radius difference

Thank you!

Q & A

www.silex-france.com