



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación



KeyCARE: A Framework for biomedical **Key**word extraction, Categorization, and semantic Relation.

Development and Use-Cases.

Sergi Marsol, BIA4NLP, Barcelona Supercomputing Center, Spain

Luis Gascó, BIA4NLP, Barcelona Supercomputing Center, Spain

Martin Krallinger, BIA4NLP, Barcelona Supercomputing Center, Spain

18/01/2024

Sergi Marsol - sergi.marsol@bsc.es

Index

1

Introduction

2

Library

1. Keyword Extractor
2. Term Categorizer
3. Relations Extractor

3

Use Cases

1. NER candidates
2. Terminology enrichment
3. Cross-ontology mapping

4

Future steps

Introduction - NLP basic concepts

- **Entity/mention** - sequence of words referring to a specific concept
- **Token** - smallest semantic unit of text in a mention
- **Keyword** - significant term or mention reflecting a document's content
- **Named Entity Recognition (NER)** - system for identifying named entities like people or places in a text
- **Corpus** - collection of documents for train-test datasets

...presenta **gastroenteritis vírica**...

mention: gastroenteritis vírica

tokens: gastro - enter - itis - vírica

A con antecedentes **SPECIES** personales de **ENFERMEDAD** diabetes mellitus tipo 2 en trat

TO encias por intensa **SINTOMA** astenia y **SINTOMA** fiebre de perfil **SPECIES** bacteriémico, **NEG** sin **DURATION** días previos.

resado en **DEPARTAMENTO** Neumología por **ENFERMEDAD** infección respiratoria, presentando ui

io servicio con la **UNC** sospecha de **USCO** fiebre de origen respiratorio, se exti
RELACION
ENFERMEDAD

Introduction

Keyword Extraction: unsupervised extraction of the important terms from a text

Term Categorization: classification of terms in semantic categories

Relations Extraction: extraction of the type of relations between terms

Keyword Extraction

Paciente ingresado con **fiebre alta** y **tos persistente**. Diagnóstico provisional de **neumonía viral**.

Se realizó **radiografía** confirmatoria y se inició **tratamiento con antibióticos** y **soporte respiratorio**.

Term Categorization

fiebre alta → SYMPTOM
neumonía viral → DISEASE
radiografía → PROCEDURE

Relations Extraction

EXACT

pirexia

NARROW

neumonía
infecciosa

BROAD

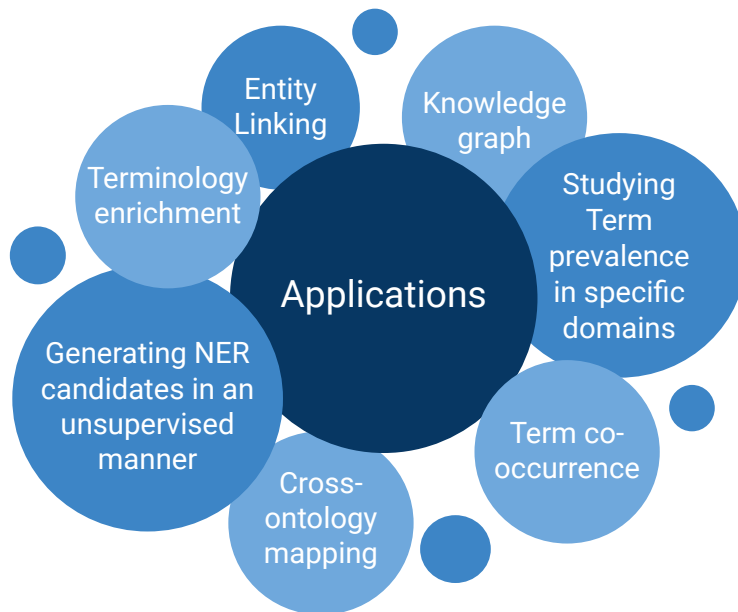
radiografía de
tórax

Introduction

Keyword Extraction: unsupervised extraction of the important terms from a text


Term Categorization: classification of terms in semantic categories

Relations Extraction: extraction of the type of relations between terms



Introduction - State of the Art

→ Named Entity Recognition (NER)

- **Supervised** systems → need for training resources [1]
- Spanish: biomedical annotated datasets by NLP4BIA 
- **Difficult scalability**, especially in low resource languages and multilingual scenarios [2]

→ Terminology Enrichment

- body of terms in a particular field of study → standard lexis and structure
- **Knowledge graphs** based on structured terminologies
- Scarce terminologies in different languages [3] → need for **semiautomatic terminology enrichment** tools [4]

KeyCARE

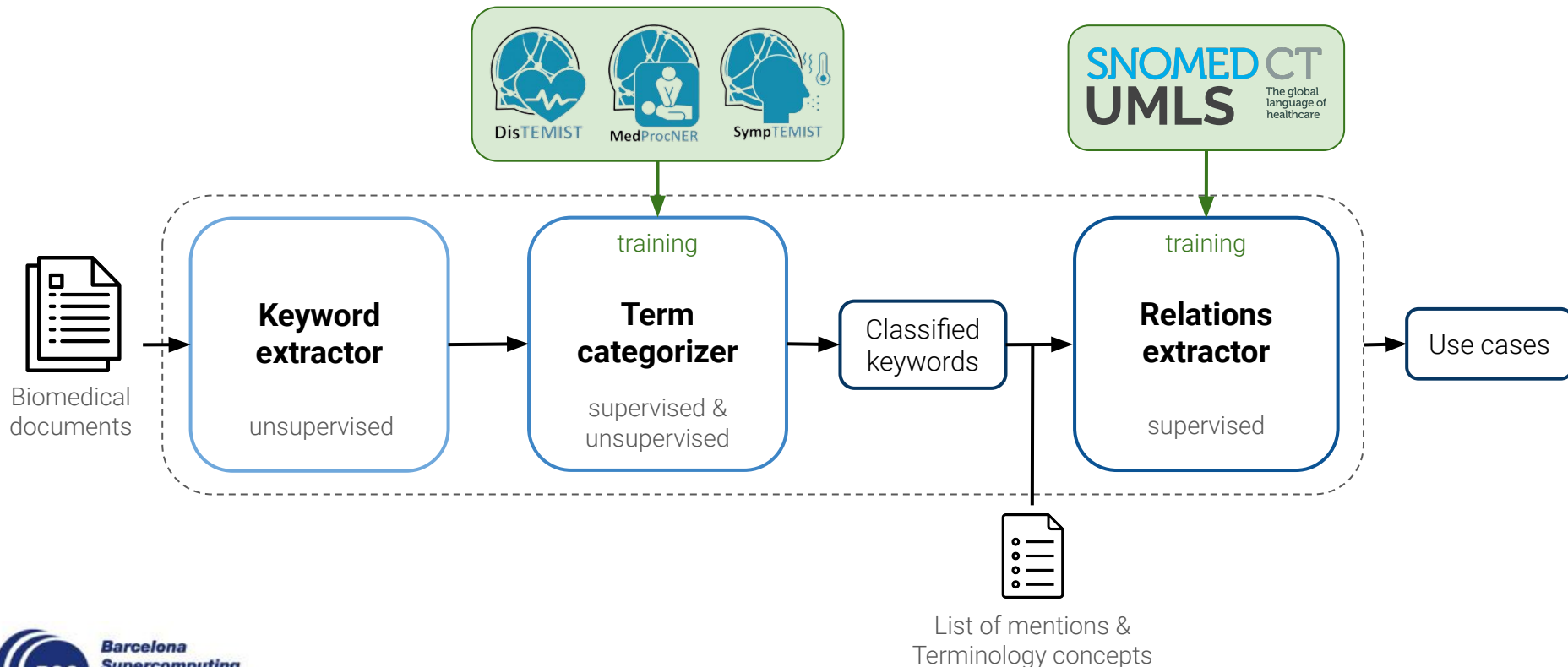
Keyword Extraction + Term
Categorization + **R**elations Extraction

Many applications in studying term prevalence and co-occurrence,
entity linking, terminology enrichment, among others.

*Soon available in Github

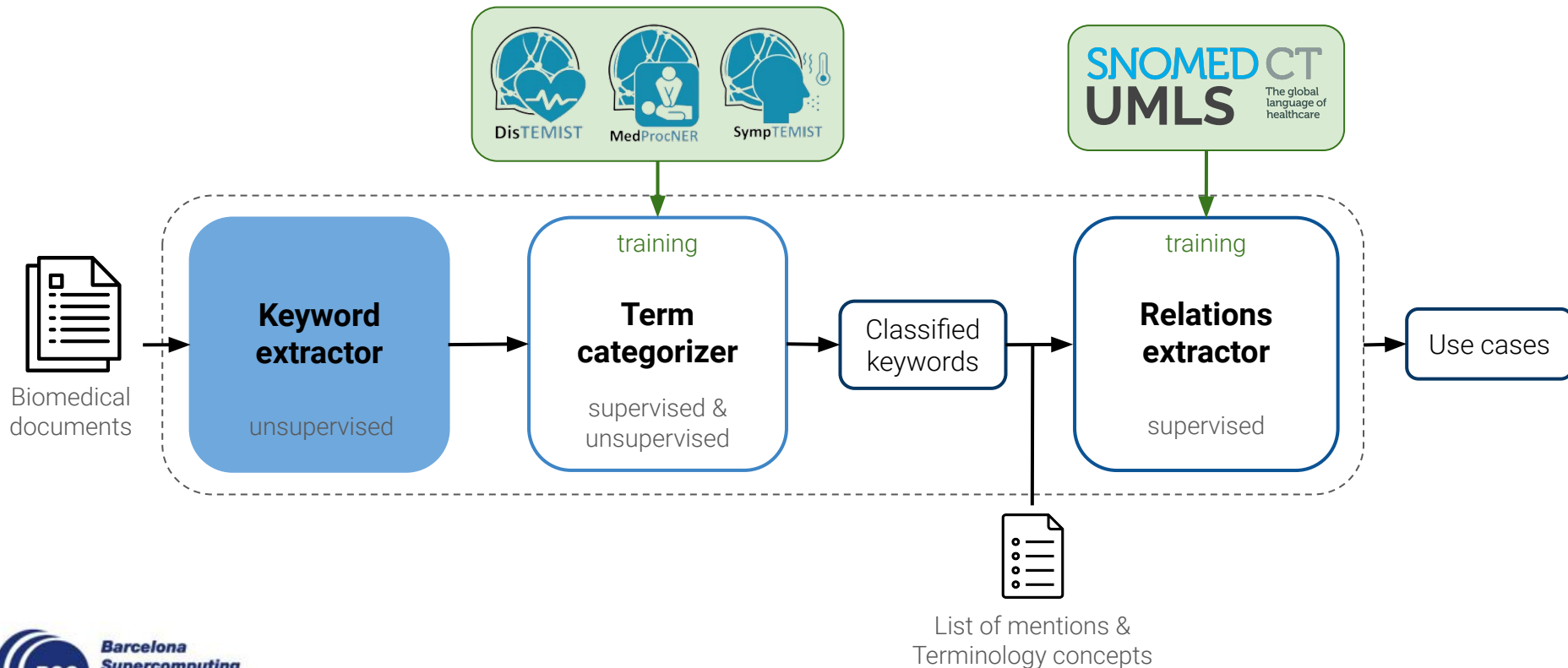
Library - General structure

Information flow —
Training data —

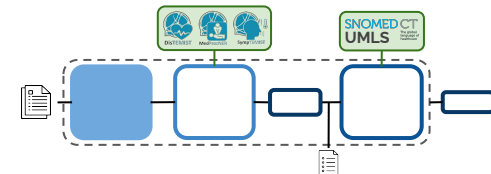


Library - Keyword Extractor

Information flow —
Training data —



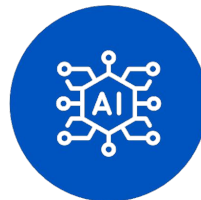
Library - Keyword extractor



Acude al Servicio de Urgencias por cefalea frontoparietal derecha.

Mediante biopsia se diagnostica adenocarcinoma de próstata Gleason 4+4=8 con metástasis óseas múltiples.

Se trata con Ácido Zoledrónico 4 mg iv/4 semanas.

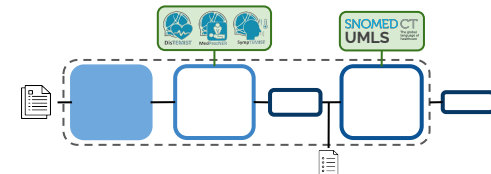


Acude al **Servicio de Urgencias** por **cefalea frontoparietal derecha**.

Mediante biopsia se diagnostica **adenocarcinoma de próstata Gleason 4+4=8** con **metástasis óseas múltiples**.

Se trata con **Ácido Zoledrónico 4 mg iv/4 semanas**.

Library - Keyword extractor



Unsupervised keyword extraction

YAKE

Extractor based on **statistical descriptors** regarding word frequency and its relationship with the context

- Language-independent
- Scalable
- Single documents

[5]

RAKE

Extractor based on the generation of a **graph of related terms** based on their **co-occurrence** to assess term importance

- Language-independent
- Domain-independent
- Single documents

[6]

TextRank

Graph-based keyword extractor based on PageRank that uses the **co-occurrence** of terms

- Language-independent
- Scalable
- Single documents

[7]

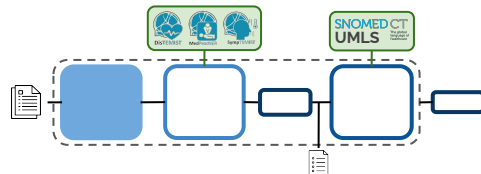
KeyBERT

Key Bidirectional Encoder Representation from Transformers based on the **semantic similarity** of words through vector representation

- **Part of Speech** tags
- Spanish SapBERT as base model

[8]

Library - Keyword extractor



Unsupervised keyword extraction

YAKE

Extractor based on **statistical descriptors** regarding word frequency and its relationship with the context

- Language-independent
- Scalable
- Single documents

[5]

RAKE

Extractor based on the generation of a **graph of related terms** based on their **co-occurrence** to assess term importance

- Language-independent
- Domain-independent
- Single documents

[6]

TextRank

Graph-based keyword extractor based on PageRank that uses the importance of terms

- Language-independent
- Single documents

[7]

KeyBERT

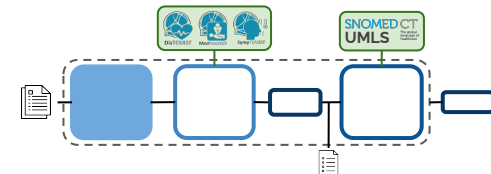
Key Bidirectional Encoder Representation from Transformers based on the **semantic similarity** of words through vector representation

- **Part of Speech** tags
- Spanish SapBERT as base model

[8]

Control of the **number of tokens** in mentions or their **PoS tags**

Library - Keyword extractor



NER Gold
Standard Corpus

MedProcNER

[9]

DisTEMIST

[10]

RAKE

precision

recall

f1-score

precision

recall

f1-score

11.93%

89.00%

21.05%

8.44%

86.41%

15.38%

YAKE

10.20%

44.72%

16.61%

9.60%

57.81%

16.47%

TextRank

13.00%

90.26%

22.73%

9.25%

88.17%

16.75%

KeyBERT

21.23%

55.18%

30.66%

15.31%

55.75%

24.02%

KeyBERT + PoS

20.22%

32.95%

25.06%

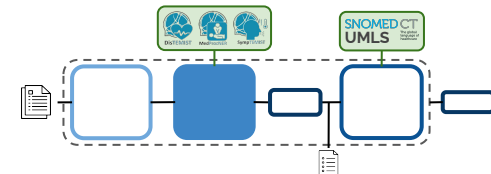
21.91%

47.87%

30.06%

*Shown scores are computed with **any degree of overlap** among terms.

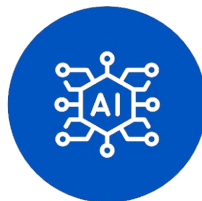
Library - Term Categorizer



Acude al **Servicio de Urgencias** por **cefalea frontoparietal derecha**.

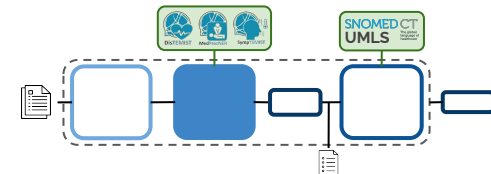
Mediante biopsia se diagnostica **adenocarcinoma de próstata Gleason 4+4=8** con **metástasis óseas múltiples**.

Se trata con **Ácido Zoledrónico 4 mg iv/4 semanas**.



Servicio de Urgencias → DEPARTMENT
cefalea frontoparietal derecha → SYMPTOM
Mediante → NO_CATEGORY
biopsia → PROCEDURE
adenocarcinoma de próstata → DISEASE
Gleason 4 → NEOPLASIA
MORPHOLOGY
metástasis óseas múltiples → DISEASE
Ácido Zoledrónico 4 → DRUG
iv/4 semanas → PROCEDURE

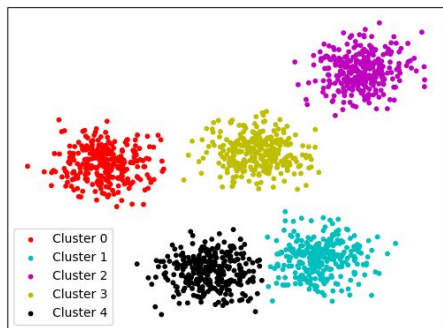
Library - Term Categorizer



Unsupervised term clustering

K-Means Clustering

Generates **clusters based on the distance** of the vectorial representations of terms



Supervised term classification

Transformers classifier

SetFit classifier

Using Spanish SapBERT
as a base model

Trained with NER Gold
Standard Corpus



MedProcNER

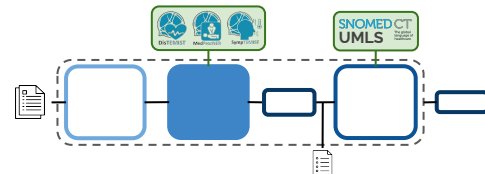


DisTEMIST



SymptTEMIST

Library - Term Categorizer classification



Using Spanish SapBERT as a base model

[11]

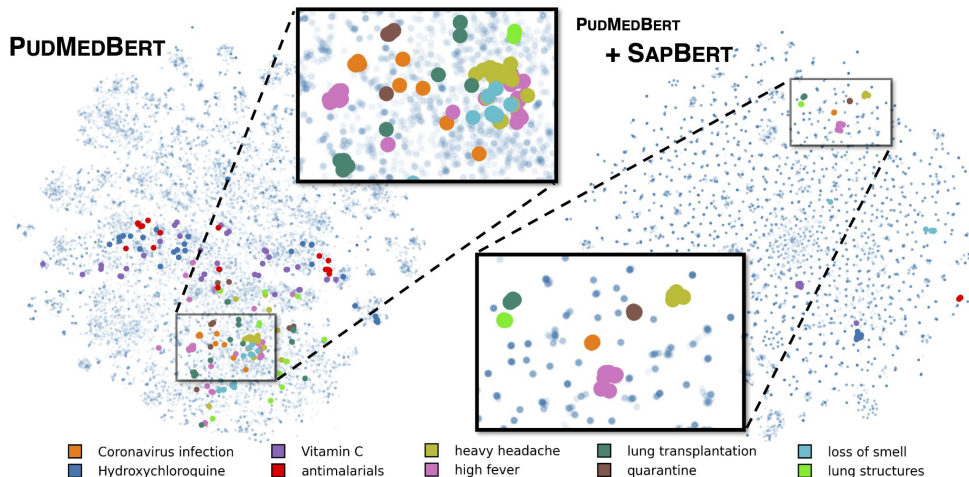
BSC-NLP4BIA / SapBERT-from-roberta-base-biomedical-clinical-es ☐ private

Feature Extraction Transformers PyTorch Spanish roberta bert biomedical lexical semantics bio

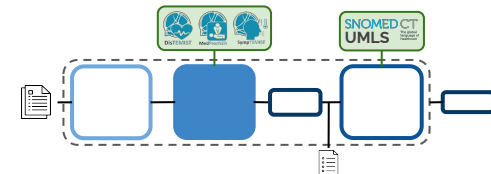
License: apache-2.0

SapBERT - Self-Alignment Pre-training for Biomedical Entity Representations

- Pretrained LM for Biomedical Entity Linking
- Using UMLS relations & contrastive learning
- SapBERT → english / multilingual / **spanish**



Library - Term Categorizer classification



Transformers classifier

Transformers **AutoModel For Sequence Classification**

- Not for Few-Shot Learning
- Fine-Tuning flexibility
- Multilingual support with base model in specified language

[12]

SetFit classifier

Few-Shot Fine-Tuning of Sentence Transformers models for sequence classification

- Fast to train
- Multilingual support with base model in specified language

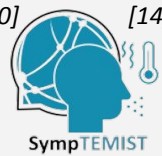
Trained with NER Gold Standard Corpus
73,863 mentions



MedProcNER



DisTEMIST

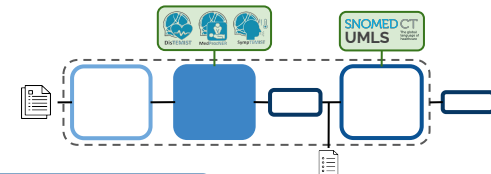


SympTEMIST

[13]

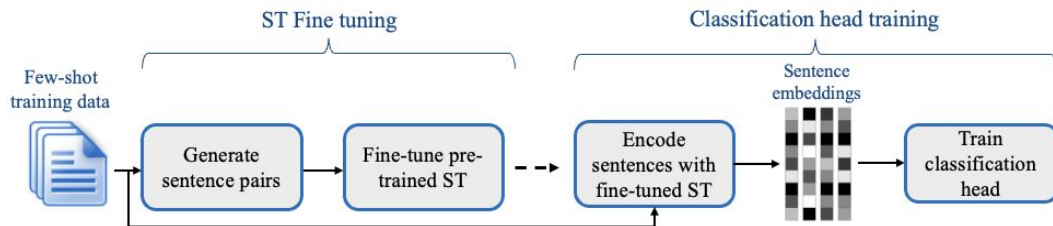
Includes 21 classes such as PROCEDURE, DISEASE, SYMPTOM, DRUG, DEPARTMENT, NO_CATEGORY, etc.

Library - Term Categorizer classification



SetFit classifier

SetFit - Sentence Transformers Fine-Tuning



Embedding Fine-tuning Phase

Contrastive Learning with sentence pairs

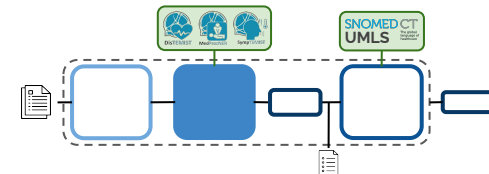
Ability to generate many unique pairs from a few examples → Few-Shot

Classifier Training Phase

Classification head for the classification of the embeddings (Logistic Regression)

[15]

Library - Term Categorizer classification



NER Gold
Standard Corpus



PROCEDURE

DISEASE

SYMPTOM

Micro avg

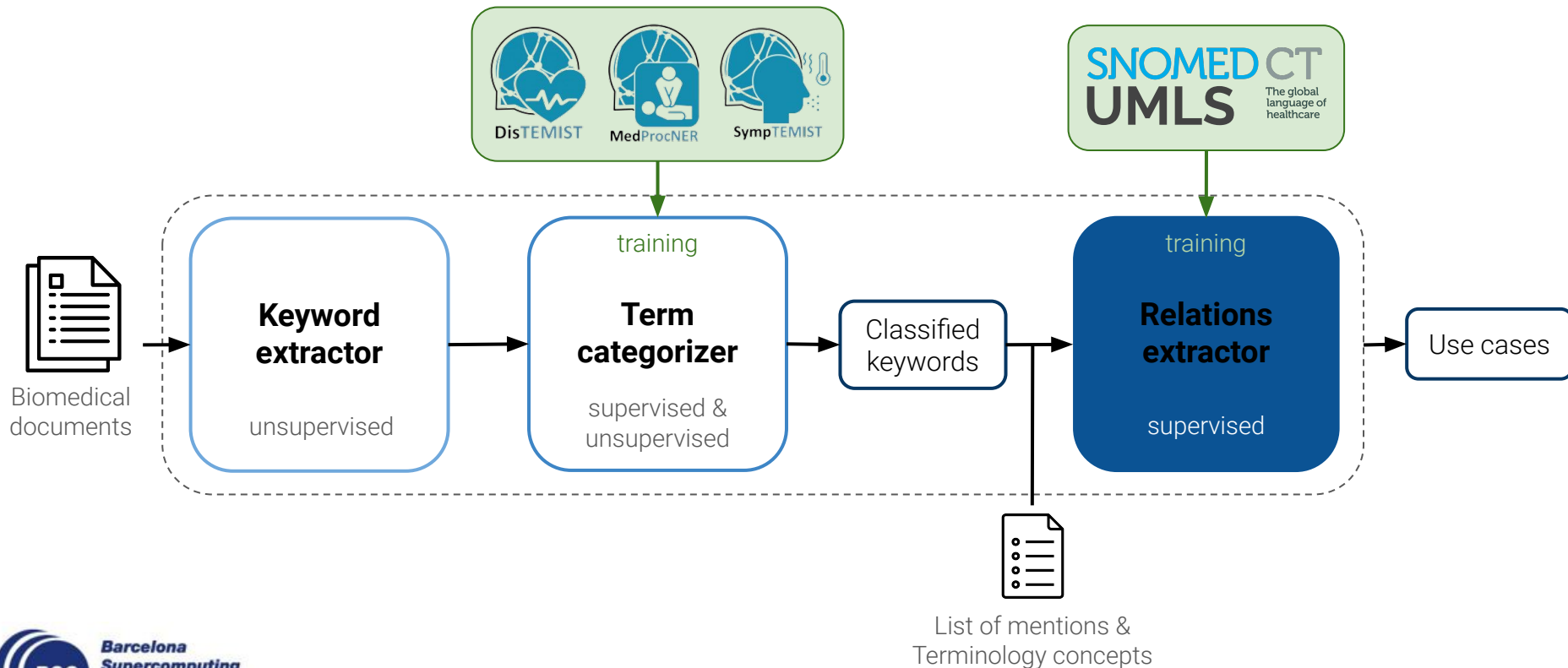
Transformers classifier

SetFit classifier

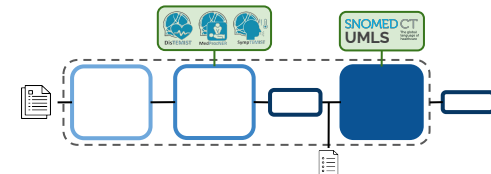
	precision	recall	f1-score	precision	recall	f1-score
PROCEDURE	96%	97%	97%	96%	97%	97%
DISEASE	84%	79%	72%	82%	82%	82%
SYMPTOM	90%	89%	89%	88%	88%	90%
Micro avg	93%	93%	93%	93%	93%	93%

Library - Relations extractor

Information flow —
Training data —



Library - Relations extractor

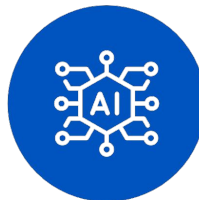


Extracted mentions (source)

metástasis óseas múltiples → DISEASE
cefalea frontoparietal derecha → SYMPTOM
biopsia → PROCEDURE
adenocarcinoma de próstata → DISEASE

Terminology concepts (target)

neoplasias malignas → DISEASE
secundarias de hueso múltiples
dolor de cabeza → SYMPTOM
biopsia de próstata transrectal → PROCEDURE
adenocarcinoma de colon → DISEASE



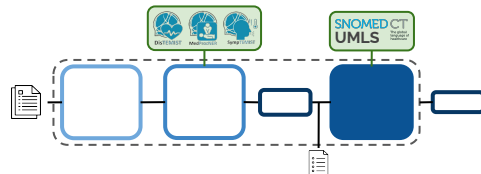
metástasis óseas múltiples
neoplasias malignas secundarias → EXACT
de hueso múltiples
Target & source are equal

cefalea frontoparietal derecha
dolor de cabeza → NARROW
Target contains source

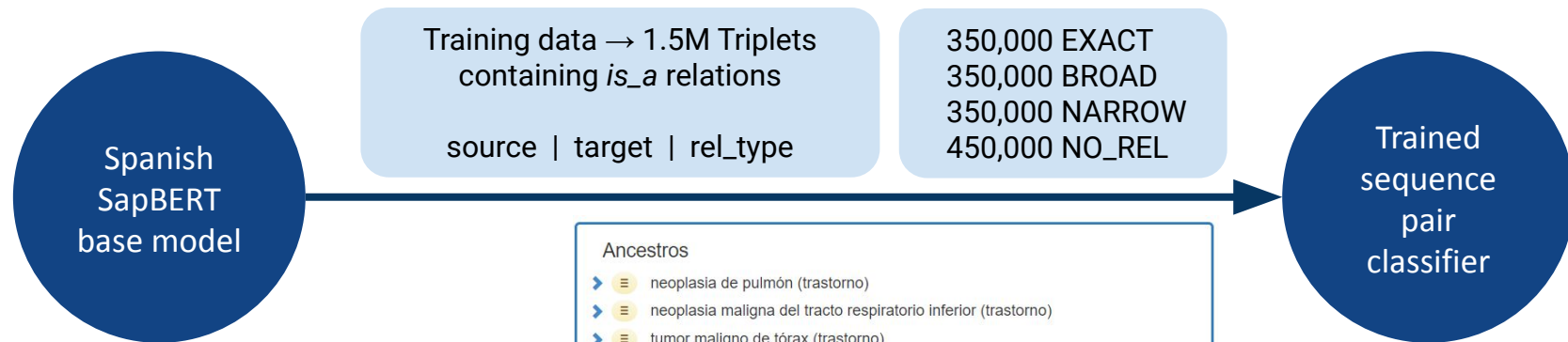
biopsia
biopsia de próstata transrectal → BROAD
Source contains target

adenocarcinoma de próstata
adenocarcinoma de colon → NO_REL
No is_a relation

Library - Relations extractor



Transformers AutoModel for Sequence Pairs Classification



SNOMED CT
UMLS
The global language of healthcare

Ancestros

- ▶ neoplasia de pulmón (trastorno)
- ▶ neoplasia maligna del tracto respiratorio inferior (trastorno)
- ▶ tumor maligno de tórax (trastorno)

tumor maligno de pulmón (trastorno)

SCTID: 363358000

363358000 | tumor maligno de pulmón (trastorno) |

- es cáncer de pulmón
- es tumor maligno de pulmón (trastorno)
- es tumor maligno de pulmón

sitio del hallazgo →
estructura del pulmón
morfología asociada →
neoplasia maligna

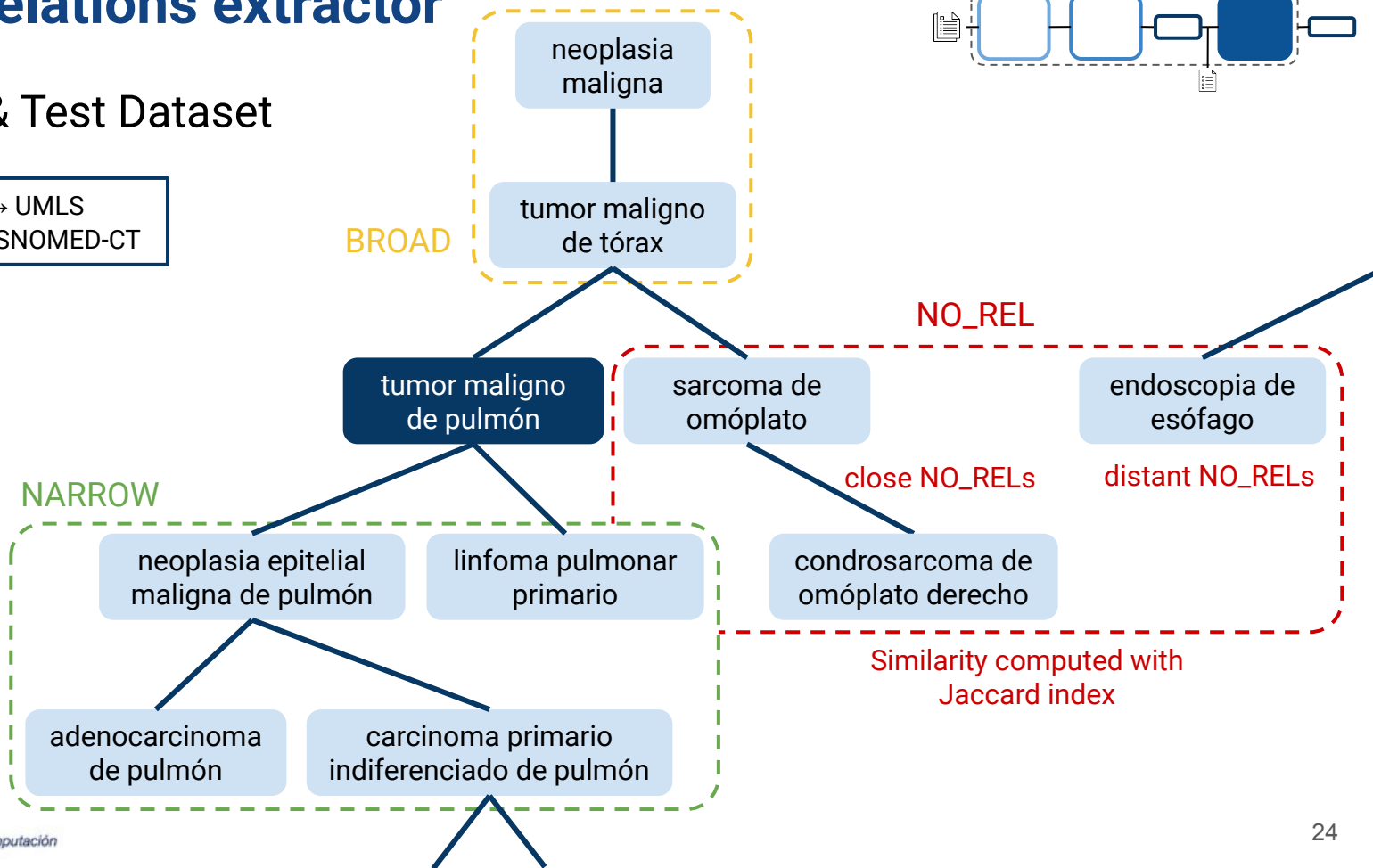
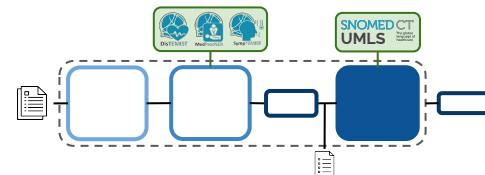
Descendientes (23)

- ▶ blastoma pleuropulmonar (trastorno)
- ▶ blastoma pulmonar (trastorno)

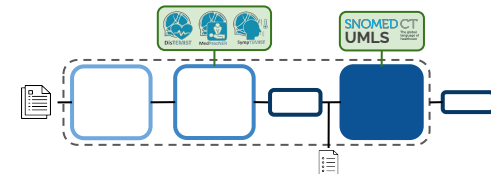
Library - Relations extractor

UMLS Train & Test Dataset

terms → UMLS
relations → SNOMED-CT



Library - Relations extractor



UMLS Test Dataset

	precision	recall	f1-score
BROAD	85%	91%	88%
EXACT	92%	93%	92%
NARROW	96%	97%	97%
NO_REL	86%	78%	82%
Weighted avg	90%	90%	90%

Manually annotated Dataset

	precision	recall	f1-score
BROAD	49%	80%	61%
EXACT	30%	94%	45%
NARROW	49%	66%	56%
NO_REL	98%	63%	77%
Weighted avg	83%	68%	71%

← External dataset with is a relations

Use cases

Specific domain example - cardiology

Introducción: De los **tumores cardíacos primarios**, más del 50 por ciento son **mixomas**. La **variabilidad sintomática** del **mixoma cardíaco**, puede llevar a **confusiones diagnósticas**.

Objetivo: Presentar un caso donde se destaca la variabilidad de **síntomas** del **mixoma cardíaco**.

Caso clínico: **Paciente masculino** de 51 años, atendido con manifestaciones de **insuficiencia cardíaca**, **trastornos** del ritmo cardíaco, **microembolias cerebrales**, **convulsiones tónico clónicas**, **hemoptisis**, **trastornos psiquiátricos**, **síndrome general** con **astenia**, **anorexia** y **pérdida de peso**. Durante dos años, fue atendido en **varias instituciones**, con **múltiples estudios** y **tratamientos**. En el **Servicio de Cardiología** de **Morón**, **Ciego de Ávila**, mediante el estudio clínico y **ecocardiográfico**, se diagnosticó un **tumor cardíaco**. Se **traslada** al **Cardiocentro** de **Santa Clara** y allí se le **extirpó** el **tumor**. Histológicamente era un **mixoma**. Evolucionó muy **bien**, con excelente calidad de vida.

Conclusiones: El **mixoma cardíaco** ocasiona **síntomas obstructivos**, **embólicos** y **constitucionales**, con cuadro clínico muy equívoco (AU).

Use cases - NER candidates

System for presenting NER candidates in a unsupervised way for low-resource languages

RAKE with up to 5 tokens	MedProcNER			DisTEMIST			
	precision	recall	f1-score	precision	recall	f1-score	
	50.26%	75.97%	60.50%	64.32%	66.84%	65.56%	
	17.23%	26.06%	20.75%	30.16%	31.35%	30.74%	Exact overlap

Use cases - NER candidates

System for presenting NER candidates in a unsupervised way for low-resource languages

	MedProcNER			DisTEMIST			
	precision	recall	f1-score	precision	recall	f1-score	
RAKE with up to 5 tokens	50.26%	75.97%	60.50%	64.32%	66.84%	65.56%	Any overlap
	17.23%	26.06%	20.75%	30.16%	31.35%	30.74%	Exact overlap
RAKE + TextRank + KeyBERT with up to 3 tokens	28.61%	94.26%	43.90%	38.45%	86.27%	53.20%	Any overlap
	12.03%	39.63%	18.46%	19.20%	43.08%	26.56%	Exact overlap

Use cases - NER candidates

System for presenting NER candidates for low-resource languages



NER Gold Standard

Entre sus antecedentes médicos destacaba padecer una **esclerosis tuberosa** y haber sido sometido a un trasplante de riñón tras sufrir una nefrectomía por un **angiomiolipoma renal**.

Presentaba como síntomas de la enfermedad el denominado **adenoma sebáceo de Pringle a nivel nasogeniano**, **hamartomas retinianos en el fondo de ojo** y **lesiones fibróticas a nivel cervical posterior**.

KeyCARE

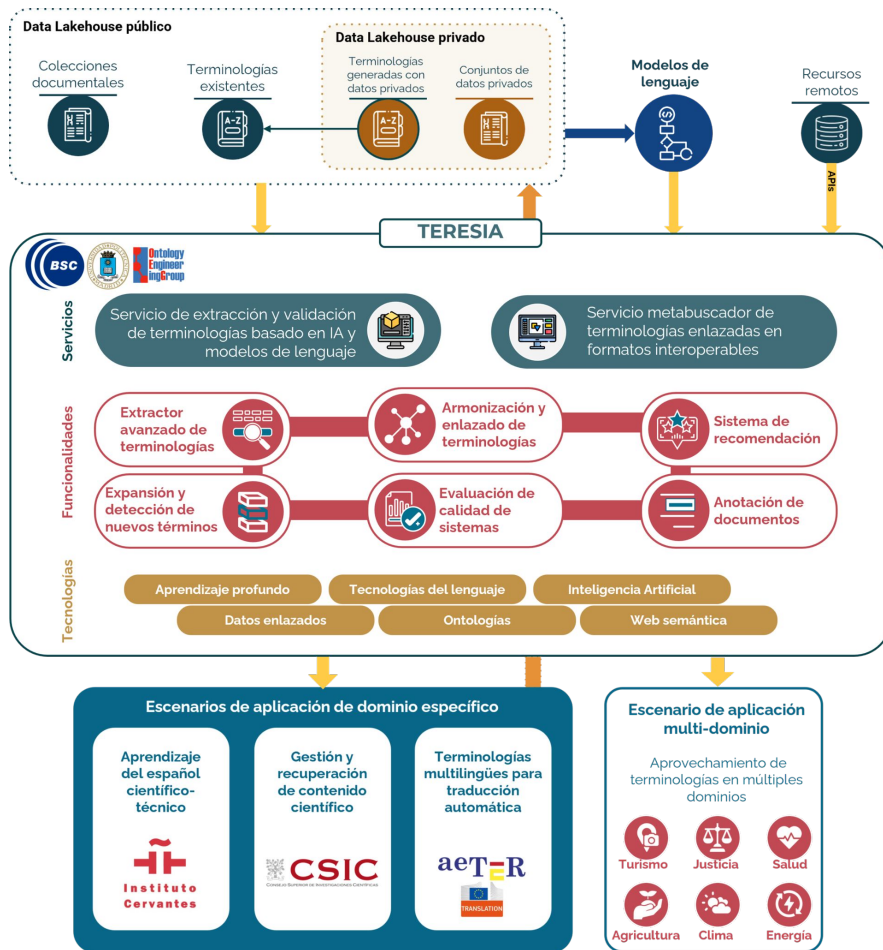
Entre sus antecedentes médicos destacaba padecer una **esclerosis tuberosa** y haber sido sometido a un trasplante de **riñón tras sufrir** una nefrectomía por un **angiomiolipoma renal**.

Presentaba como síntomas de la **enfermedad** el **denominado adenoma sebáceo** de Pringle a nivel nasogeniano, **hamartomas retinianos** en el fondo de ojo y **lesiones fibróticas** a nivel cervical posterior.

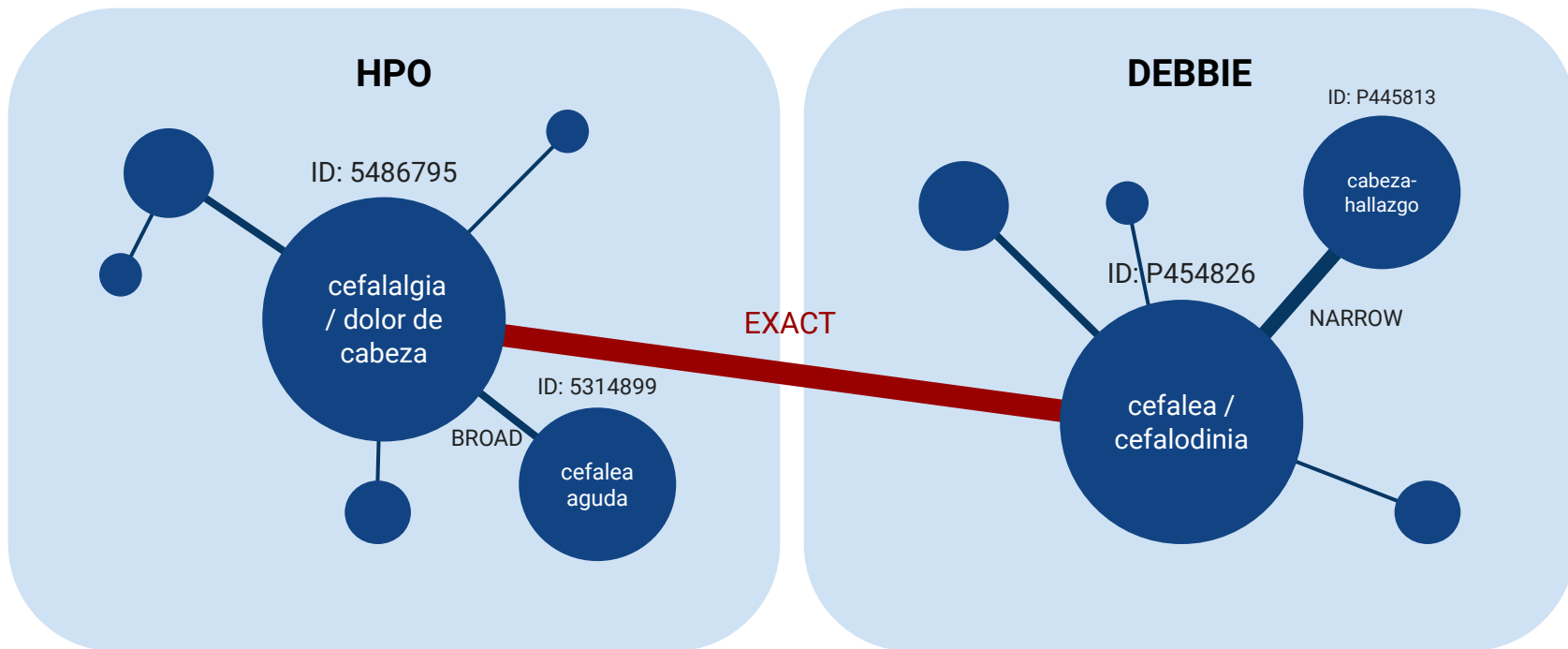
Use cases - Terminological enrichment

Enrichment of relations between terms & discovery of new terms



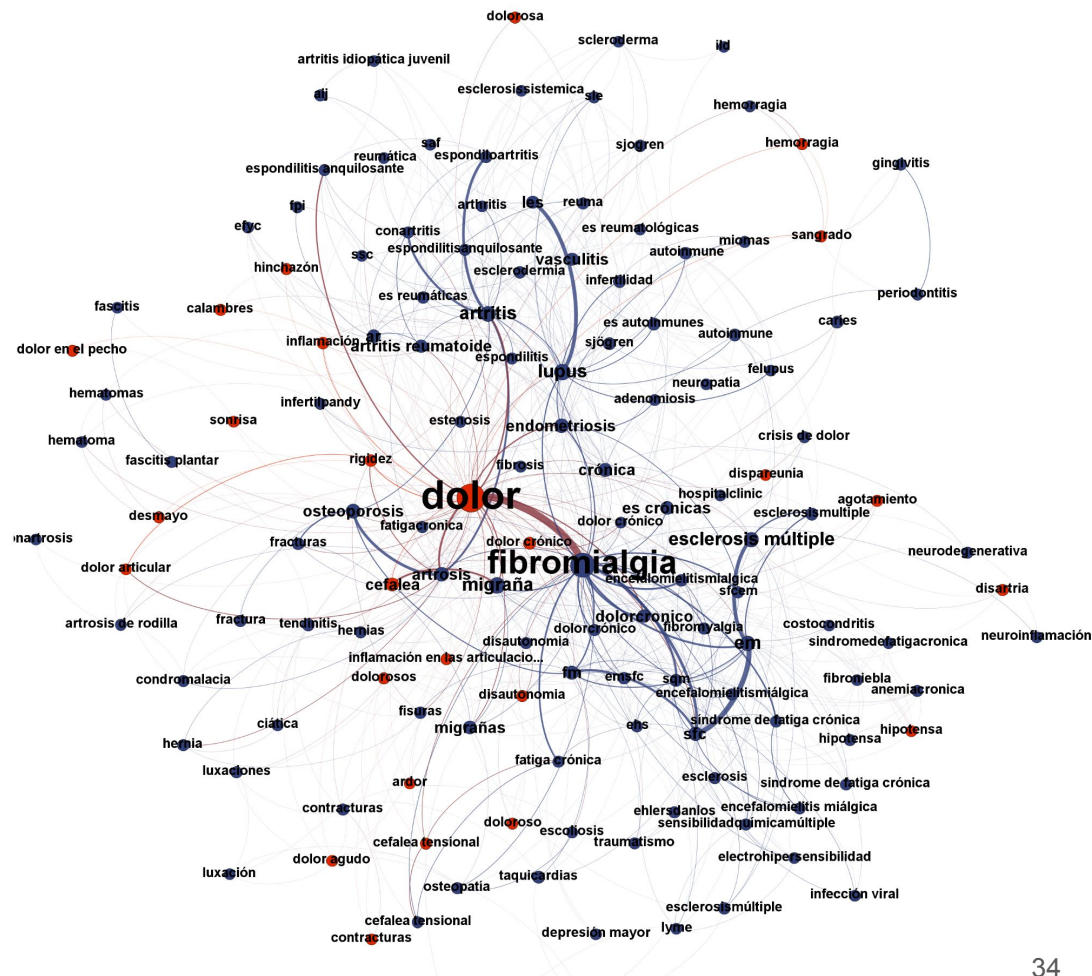


Use cases - Cross-ontology mapping



Use cases - Others

- Knowledge graph
- Analysis of term prevalence in specific domains and of term co-occurrence
- Entity Linking improvement



Future steps

The Use Cases are not yet implemented but some are already doable:

NER candidates - through evaluation in other languages

Terminology enrichment - through the development of an interface

Analysis of **term prevalence in specific domains** and of **term co-occurrence**

...

KeyCARE

A python library for biomedical keyword extraction, term categorization, and semantic relation

[Explore the docs »](#)

[Report Bug](#) · [Request Feature](#)



Table of Contents

1. [About the Project](#)
2. [Getting Started](#)
 - 2.1. [Installation](#)
 - 2.2. [Usage](#)
3. [Contributing](#)
4. [License](#)
5. [References](#)



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación

Thank you!

18/01/2024

Sergi Marsol - sergi.marsol@bsc.es