Text mining processing pipeline for semi structured data D3.3

Copara, Jenny; Naderi, Nona; Kellmann, Alexander; Gosal, Gurinder; Hsiao, William; Teodoro, Douglas

DOI:
10.5281/ZENODO.5795433

Common Infrastructure for National Cohorts in Europe, Canada, and Africa
- CINECA -

D3.3 - Text mining processing pipeline for semi structured data

Work Package:	WP3 - Cohort Level metadata Representation
Lead Beneficiary:	European Molecular Biology Laboratory
WP Leaders:	Fiona Brinkman (SFU), Melanie Courtot (EMBL-EBI)
Contributing Partner(s):	UMCG, SFU, UCT, HES-SO, SIB, EMBL-EBI
Contractual Delivery Date:	31st December, 2021
Actual Delivery Date:	16th December 2021
Authors of this Deliverable:	Jenny Copara (SIB), Nona Naderi (SIB), Alexander Kellmann (UMCG), Gurinder Gosal (SFU), William Hsiao (SFU), Douglas Teodoro (SIB)
Contributors:	Isuru Liyanage (EMBL-EBI)
Reviewed by:	Jonathan Dursi/Jordi Rambla
Approved by:	Thomas Keane (EMBL-EBI)
Dissemination Level:	Public
Type of Deliverable:	Other
Grant agreement:	No. 825775 Horizon 2020 (H2020-SC1-BHC-2018-2020)
Type of action:	RIA
Start Date:	1 Jan 2019
Duration:	48 months

Table of contents:

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

1. Executive Summary

Unstructured and semi-structured cohort data contain relevant information about the health condition of a patient, e.g., free text describing disease diagnoses, drugs, medication reasons, which are often not available in structured formats. One of the challenges posed by medical free texts is that there can be several ways of mentioning a concept. For example, one may use the passage "heart attack", "myocardial infarction", or "MI" to refer to the same medical concept. Encoding free text into unambiguous descriptors allows us to leverage the value of the cohort data, in particular, by facilitating its findability and interoperability across cohorts in the project. Normalization of free text also contributes to populating the minimal metadata model of cohort free text fields.

Named entity recognition and normalization enable the automatic conversion of free text into standard medical concepts. Given the volume of available data shared in the CINECA project, the WP3 text mining working group has developed named entity normalization techniques to obtain standard concepts from unstructured and semi-structured fields available in the cohorts. Independent and targeted normalization solutions were designed for specific cohorts after identifying normalization needs. For example, in the CHILD cohort, disease, drug name, and medication reason are the fields requiring cleaning and normalization, while in the CoLaus/PsyCoLaus cohort, several fields associated with diagnosis require normalization. Then, a final text mining aggregator interface was developed to integrate the different solutions in a common framework.

In this deliverable, we present the methodology used to develop the different text mining tools created by the dedicated SFU, UMCG, EBI, and HES-SO/SIB groups for specific CINECA cohorts. Individual solutions were deployed locally to avoid moving sensitive data to an external environment. LexMapr, developed by SFU, obtained an accuracy of 0.88 and F1-measure of 0.83 for the normalization of the CHILD cohort test dataset, also providing a recall of 0.78. SORTA, developed by UMCG, achieved a precision/recall of 0.58/0.40 for mapping Dutch phenotypes to HPO and in a second use case 0.59/0.65 for mapping free text physical activity to MET (metabolic equivalent task) ontology. ZOOMA, developed by EBI, achieved precision/recall of 0.63/0.17 for the same experiment as the SORTA for mapping Dutch phenotypes to HPO. L2N, developed by HES-SO/SIB, achieved an accuracy of 0.76 in the Medical Concept Normalization corpus. L2N was applied to the diagnoses fields of the CoLaus/PsyCoLaus synthetic data. The main text mining API, integrating these different services, is deployed as a web service that encapsulates and exposes the pipeline functionality of each group.

To populate the minimal metadata model, the text mining working group has developed different tools to assign standard medical concepts to free text in CINECA cohorts. The tools were evaluated on different de-identified benchmark datasets, achieving varied performance levels depending on the dataset and task complexity, but compatible with the state-of-the-art. The resulting models were then applied to the project's cohort free text data according to their specific normalization needs. Even though each group developed independent solutions, an integrated API was deployed containing the functionality of each tool. Our next

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

step is the alignment to the Genomics Cohorts Knowledge Ontology (GECKO). We also will continue refining our pipelines to get more accurate normalization results in cohort data. Finally, as we have different pipelines to normalize concepts, we aim to create a common test set to assess the performance of the normalization methods.

2. Project objectives

WP3 Task 3.4 objective:

1. To develop and apply text mining strategies for the population of the minimal meta data model where data are unstructured or semi structured.

3. Detailed report on the deliverable

3.1 Background

CINECA cohorts were developed under different contexts to cover specific goals, targeting population groups by age, health condition, diseases, etc. Cohort data are shared in a variety of formats and modalities, bringing a valuable source of information for further research. A key characteristic of these rich cohorts is that data are heterogeneous in terms of format and content. Combining different cohorts demands to define a minimum set of cohort fields that are more relevant to synthesize each cohort and to allow the search across cohorts, i.e., the minimum metadata model developed in WP3. To enable search and knowledge discovery across the heterogeneity of several cohorts, we need to represent the cohort data using a common, and unified language.

Named entity recognition and normalization extract automatically standardized terms from free text data [1]. There are different tools to annotate concepts, rule-based, machine learning based, or a combination of both. For example, MetaMap, a machine learning and rule-based tool developed by the US National Library of Medicine, maps biomedical text to UMLS identifiers [2]. MetaMap performs a lexical and syntactic analysis consisting of tokenization, part-of-speech tagging, lexical lookup of input words, word variant generation, and word sense disambiguation. The clinical text analysis and knowledge extraction system (cTAKES) is trained on the clinical domain data to extract clinical information from unstructured text. Its components include sentence boundary detector, rule-base and context-dependent tokenizer, part-of-speech tagger, phrasal chunker, negation detector, dependency parser, and drug mention annotator. cTAKES extracts concepts from SNOMED-CT, UMLS, and RxNorm [3].

To develop methods for automatic concept annotation, it is fundamental to have manually annotated datasets, or gold-standard, in the medical domain. Annotated data is useful to learn to identify entities in this domain, i.e., concepts in medical ontologies/terminologies. For example, the Medical Concept Normalization (MCN) corpus [1] contains discharge summaries annotated with UMLS, a meta-ontology covering and integrating NCI Thesaurus, ICD-10, and HPO concepts, among others. It was used, among others, in the National NLP Clinical Challenges (N2C2) shared task on clinical concept normalization.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

Using the MCN corpus and the context of the discharge summaries, Chen et al. (2020) developed a hybrid machine learning model combining dictionary lookup, contextualized word representations, and the Siamese attention architecture obtaining 82.09% of accuracy in MCN [4]. Using an exact match approach (against the MCN training set and UMLS) and MetaMap, Luo et al. (2019) obtained 75.65% accuracy [1]. The system performing the best accuracy in the challenge achieved 85.26%. It was trained with the MCN training set and UMLS. This system used a contextualized word representation trained on scientific texts [5].

In the context of CINECA cohorts, free text passages usually come from tabular data, and thus lack context and rather contain few words describing the health condition, drugs, or medication. Moreover, as we are dealing with sensitive data, solutions need to be deployed locally for specific cohorts. Given these constraints, the WP3 text mining working group addressed the concept normalization needs over free text cohort data focusing on the following objectives:

1. To develop or adapt existing text mining solutions for concept normalization of free text.
2. To validate the performance of the concept normalization pipeline.
3. To apply the concept normalization pipeline to the target cohort.
4. To integrate the different pipelines into a web service to allow queries of free text or input files.

Independent pipelines have been developed targeting different cohorts. To better address each text mining cohort's needs, fields containing free text were identified. Each specific text mining tool provides a web service endpoint to allow API integration. The functionality of these pipelines is made available through an integrated text mining API.

3.2 Description of Work

In this section, we describe the CINECA cohort data sets, the relevant information about the developed synthetic cohort data sets during the CINECA project, the methods developed or extended by each working group and their validation results, the integrated text mining API built, and finally we demonstrate the application the concept normalization tools on selected synthetic data sets.

3.2.1 Data

1. **CoLaus/PsyColaus**
   CoLaus is a population-based study of 6,188 subjects from Lausanne, Switzerland [6]. It studies the prevalence of cardiovascular risk factors and the genetic determinants associated with cardiovascular risk factors. Some fields in this cohort contain free text to describe personal or family history of the disease. PsyCoLaus is the psychiatric branch of CoLaus study, in which a total of 3,691 individuals participated [7]. PsyCoLaus studies the prevalence of psychiatric syndromes. It also studies associations between psychiatric disorders, personality traits, and cardiovascular diseases. Finally, PsyCoLaus studies the genetic variants that affect

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

the risk for psychiatric disorders and whether genetic risk factors are shared between psychiatric disorders and cardiovascular diseases. Both cohorts share data within the CINECA project.

To speed up development and minimise data sharing, a synthetic dataset based on the CoLaus/PsyColaus content and fields was created[1]. The synthetic CoLaus/PsyCoLaus data considers two fields of free text from the cohort, i.e., dginvtx2 and dginvtx3, which contain disease diagnoses. These fields are the most populated among all free text fields in this cohort. Though these fields are dedicated to disease diagnoses, they also contain procedure, symptom, finding, and drug information.

2. **CHILD**

The CHILD Cohort Study[2] is a multidisciplinary, longitudinal, population-based birth cohort study in Canada that studies how various early-life exposures link to health and disease outcomes. In this cohort study, the health of approximately 3,500 Canadian children has been tracked by researchers to discover means to avert asthma, allergies, obesity, and other chronic diseases. Synthetic CHILD cohort dataset (CINECA synthetic cohort NA Canada CHILD[3]) was developed to describe how data is structured for select common attributes in the CHILD Cohort Study without revealing any individual or identifiable information associated with cohort participants. It comprises 100 variables for synthetic participants who have faked phenotypic data that reflects CHILD cohort data. In addition, there is genetic data based on the 1000 Genomes project. CHILD cohort data contains a few fields, such as medication name and medication reason, that describe the values in free text. These free-text fields have been the focus of text-mining tasks for normalization and standardization.

3.2.2 Methods

In this section, we describe the four concept normalization pipelines developed by the WP3 text mining working group for processing semi structured data: ZOOMA, SORTA, LexMapr, and L2N. ZOOMA is developed by EBI, while SORTA is developed by UMCG. LexMapr has been developed by SFU. Finally, HES-SO/SIB has developed the pipeline L2N.

1. **ZOOMA**

ZOOMA is an ontology annotation tool developed at EMBL-EBI for mapping free text to ontology terms. ZOOMA is backed by a linked data repository of annotation knowledge which contains curated annotations derived from many publicly available data sources such as Expression Atlas, Open Targets and GWAS Catalog. Therefore

[1] https://doi.org/10.5281/zenodo.5082689
[2] https://childstudy.ca/
[3] https://zenodo.org/record/5122832#.YZBrFGBKg2x

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

ZOOMA can facilitate annotations relating to a diverse range of topics, including disease and phenotypes, drug treatments, anatomical components, species, cell types, etc. Furthermore, ZOOMA can be easily configured to use new data sources or prioritise certain data sources over others to enhance context sensitivity. To increase FAIRness [8] of its data, EBI BioSamples has developed a pipeline to automatically annotate sample attribute-values with ontologies using ZOOMA. This allows researchers to do complex queries using ontology expansion and synonyms - for example, searching for heart diseases will return samples annotated with myocardial infarction using ontology expansion. Figure 1 depicts a sample label and value annotated with ZOOMA.

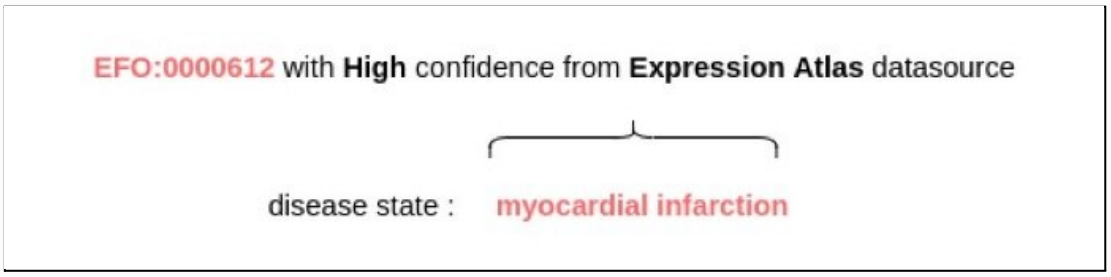

Figure 1. Label-value annotated with ZOOMA including confidence and the annotated ontology.

2. SORTA

UMCG has developed SORTA, a tool to recode free text answers with standard Ontology terms. SORTA stands for a System for Ontology-based Re-coding and Technical Annotation of biomedical phenotype data. In contrast to the other tools, SORTA is a semi-automatic tool. It provides a list of the best matching suggestions, while a user needs to take the final decision. This makes it possible to deal with misspelt terms. The user can choose to accept suggestions above a given score threshold automatically. An overview of SORTA's strategy is shown in Figure 2. SORTA combines a lexical-based and a semantic matching approach to do the matching and to generate a score. It can also take synonyms into account. SORTA uses Apache Lucene, a high-performance search engine, which uses a token-based algorithm. The Lucene matching scores are not comparable across different queries, which makes it unsuitable for human evaluation. Therefore the n-gram based algorithm was added as a second matcher. It allows to standardize the similarity scores as percentages (0-100%) which helps users to understand the quality of the match and to enable a uniform cut-off value [9]. Finally, the system allows users to upload and choose a domain-specific Ontology to SORTA and to perform a mapping between a list of terms and one of the available ontologies.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

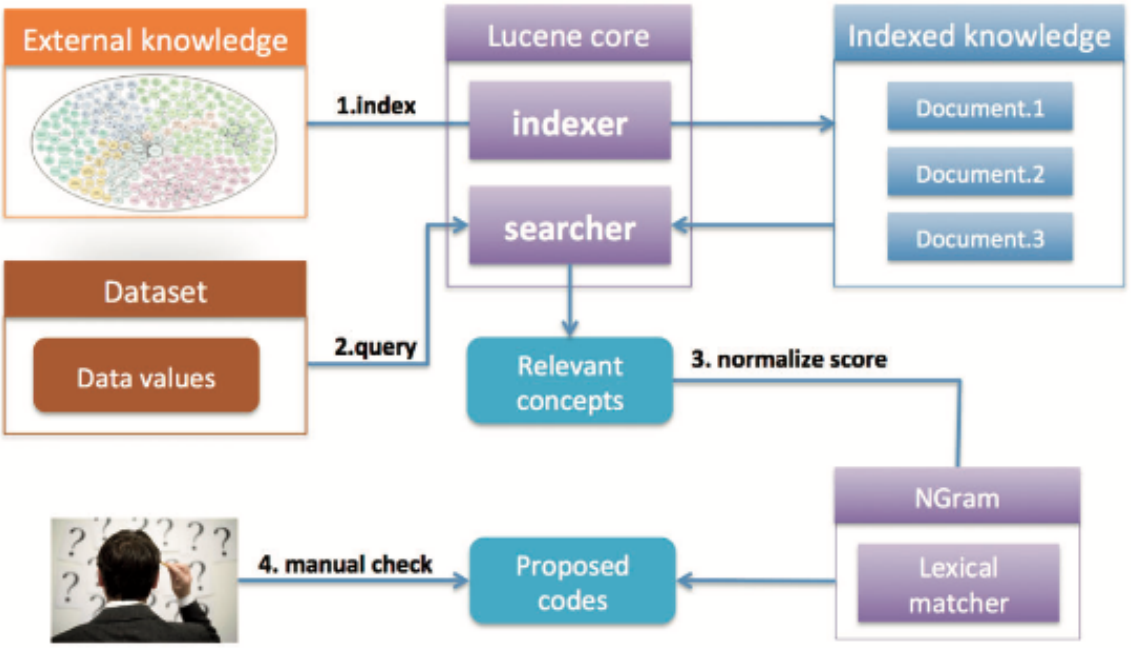

Figure 2. SORTA overview, originally from [9].

3. LexMapr

SFU (Hsiao Lab) has developed a rule-based text-mining tool called LexMapr that cleans up and parses unstructured text in the form of short phrases to extract biomedical entities and links these to standard ontology terms. Although LexMapr was initially developed to serve the biosample domain, LexMapr's general approach of cleaning and harmonization of data allowed it to be adapted and used to address different domains by adding selected domain-specific ontologies and rules. To complement other text mining tools in WP3 in the CINECA project, LexMapr provides cleaning, normalization, and ontology term linking methods by focussing on the narrative contents i.e. free-text field values of cohort data. It combines basic lexicographic transformation with Natural Language Processing (NLP), synonymy, ontology, and other resource lexicons to accomplish these tasks. The LexMapr pipeline addresses many challenges in the processing of short biomedical phrases in cohort data such as grammatical inadequacies (inconsistent use of letter cases and punctuation), spelling mistakes, the use of non-standard abbreviations, overlapping biological entities, the arbitrary ordering of words, and the inconsistencies in the ontology term labels.

LexMapr Pipeline

The initial focus of LexMapr development has been on providing a text-mining tool to clean up the short free-text biosample metadata that contained inconsistent punctuation, abbreviations and typos, perform synonym and abbreviation normalization, and to link the identified entities to standard terms from ontologies. Because the problem space of short phrases pose very specific challenges, LexMapr

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

has used a rule-based approach that draws upon wide-ranging lexical resources. LexMapr implements different rules for pre-processing, normalization, entity recognition and ontology term mapping tasks, and makes use of domain-specific customized lexicons for abbreviation and acronyms normalization, usage of trademark (brand) names, and a controlled use of stopwords elimination and singularization. LexMapr uses a specialized spell corrector tool BioSpellC developed in the Hsiao Lab (SFU) for correcting the spelling mistakes in the short phrases.

LexMapr pre-processes the input short descriptions by implementing a series of steps for data cleaning, punctuation and case treatment, singularization, and spelling correction. The pre-processing phase improves the results by providing cleaned phrases for subsequent steps of entity recognition and term mapping by LexMapr. The normalization phase transforms the entities to their normalized forms before term mapping is performed. LexMapr normalizes the usage of abbreviations or acronyms in input descriptions obtained from the previous phase. In the term mapping phase, LexMapr makes use of several rules on pre-processed and normalized samples to support the detection of relevant entities and map to ontology terms. Figure 3 shows the high-level architecture of LexMapr and its different enabling components.

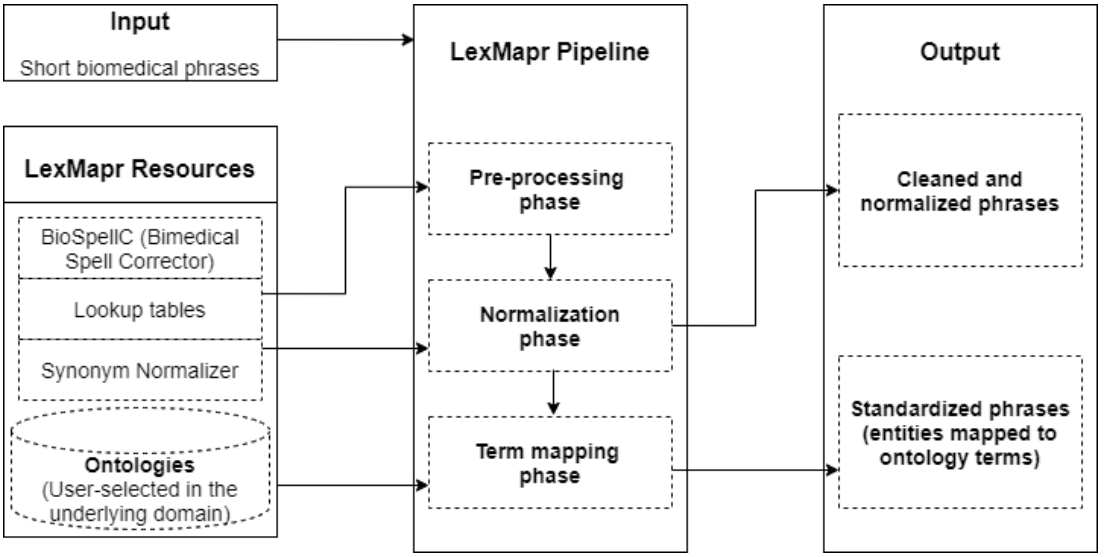

Figure 3. LexMapr's general architecture.

For the initial experimentation, the Table 1 list of OBO Foundry [10] ontologies have been selected as the target ontologies for standardizing input text in disease and drug domains. LexMapr is designed to allow customized selection of ontologies and lexical resources, therefore, the user could easily configure the set of ontologies or branches of ontologies as target ontologies for standardization.

Ontologies/ontology branches	**Resource link**

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

MONDO	http://purl.obolibrary.org/obo/mondo.owl
DOID	http://purl.obolibrary.org/obo/doid.owl (branch-http://purl.obolibrary.org/obo/DOID_4)
CIDO	http://purl.obolibrary.org/obo/cido.owl
IDO	http://purl.obolibrary.org/obo/ido.owl
HTN	https://raw.githubusercontent.com/aellenhicks/htn_owl/master/htn.owl
MFOMD	http://purl.obolibrary.org/obo/MFOMD.owl
SCDO	http://purl.obolibrary.org/obo/scdo.owl
HP	http://purl.obolibrary.org/obo/hp.owl (branch - http://purl.obolibrary.org/obo/HP_0000001)
SYMP	http://purl.obolibrary.org/obo/symp.owl
CMO	./cinecamapr/lexmapr/tests/test_ontologies/cmo.owl (locally stored)
GECKO	http://purl.obolibrary.org/obo/gecko/ihcc-gecko.owl
PDRO	http://purl.obolibrary.org/obo/pdro.owl (branch - http://purl.obolibrary.org/obo/BFO_0000040)
OMIT	http://purl.obolibrary.org/obo/omit.owl
DRON	http://purl.obolibrary.org/obo/dron.owl (branch-http://purl.obolibrary.org/obo/BFO_0000001)
CHEBI	http://purl.obolibrary.org/obo/CHEBI.owl (branch - http://purl.obolibrary.org/obo/CHEBI_24431)

Table 1. Set of ontologies selected in disease and drug domains for standardizing input text.

The different rules have been implemented to deal with the irregular case usage, long names, naming variations, and word ordering in input phrases and ontology term labels and suffix addition to input text. In case of no direct mention of the entities in the input phrases, LexMapr attempts to map entities to standard ontology terms (indirectly) by making use of synonyms. For synonym substitution, LexMapr primarily makes use of the exact synonyms for standard terms available in the selected ontologies. Also for the CINECA-customized version, LexMapr makes use of an OWL (Web Ontology Language) file that is a placeholder of different trademark names used in the underlying domains of disease and drug. The mapped set of terms are further refined with the ontology-driven pruning (using the hierarchical structure of the

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

ontologies) to retain more specific terms when multiple mappings are obtained. Figure 4 shows the different phases of the LexMapr pipeline.

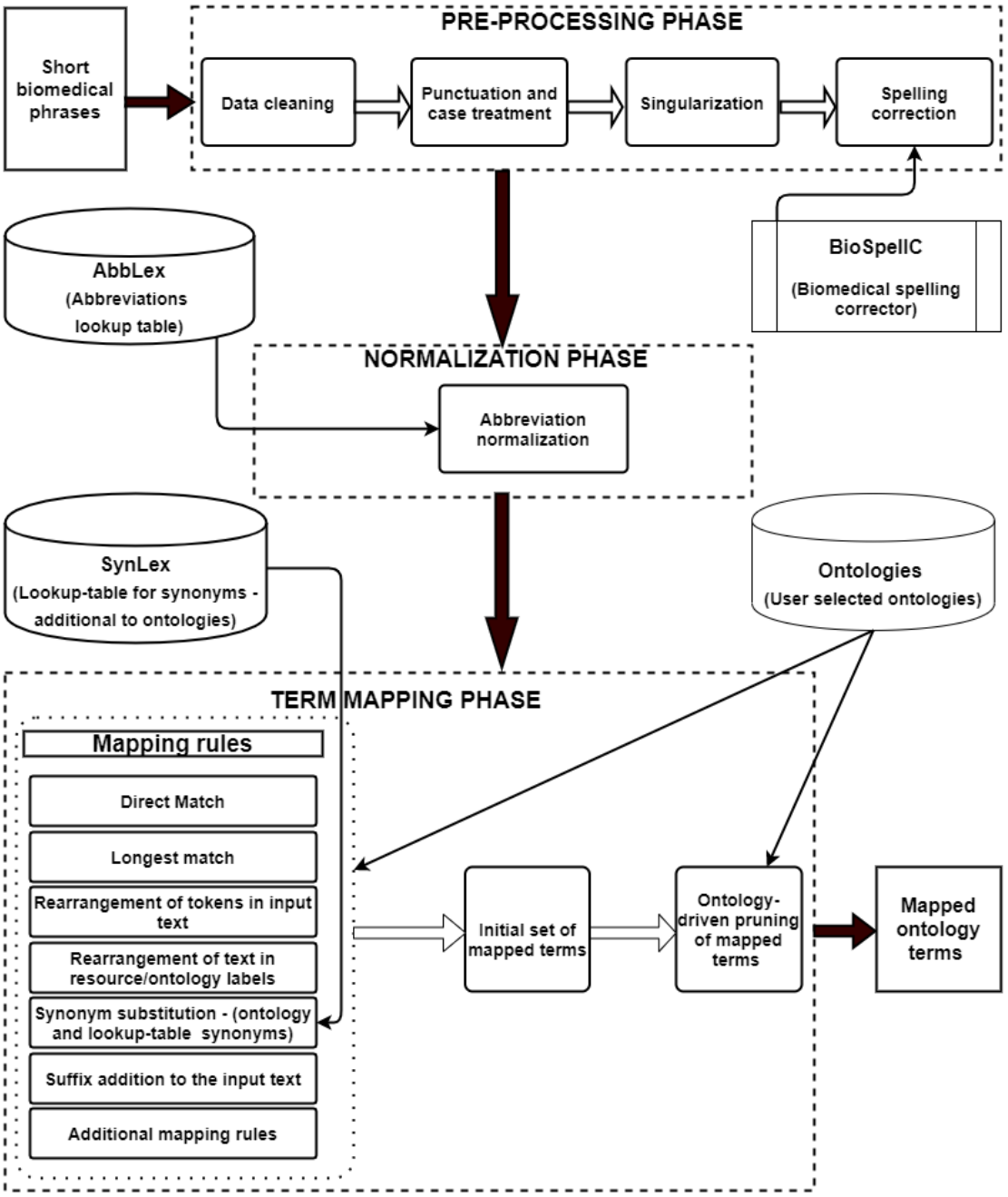

Figure 4. LexMapr text-mining pipeline.

4. L2N

HES-SO/SIB has developed a normalization pipeline for biomedical free texts called Learning to Normalize (L2N). Differently from previous methodologies, it uses

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

machine learning methods in the normalization process. After the rule-based pipeline, supported by MetaMap, it applies methods of learning-to-rank to improve the score of the candidate terms. Figure 5 shows the normalization pipeline. The system receives a free-text passage as input and provides a normalized concept as output using the Concept Unique Identifier (CUI) ids of the UMLS Metathesaurus. The normalization is performed through a dictionary matching module, a learning to rank module, and a spelling corrector. First, the input text goes to the dictionary matching module, through the exact match against manually annotated data, e.g., the MCN corpus and UMLS. The next step is a query to Metamap [2]. The dictionary matching module aims to map the input data to exactly one concept. In the case of more than one MetaMap result, the learning to rank module is applied.

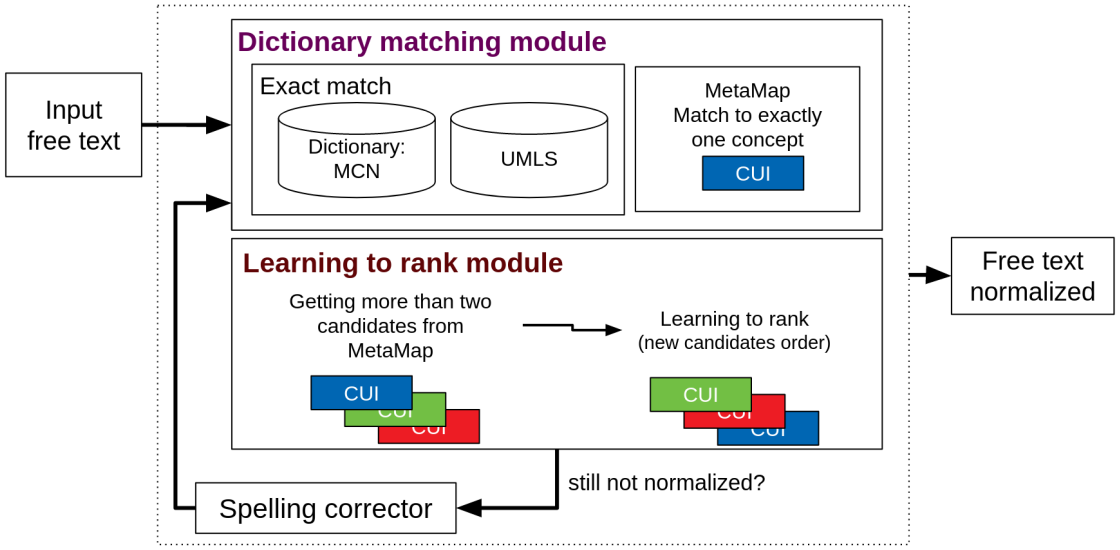

Figure 5. L2N normalization pipeline.

The learning-to-rank algorithm learns to reorder the ranked list provided by the dictionary matching module. MetaMap provides an initial list of candidates for the normalization. Then, the learning to rank module re-ranks the candidates to get a new order of concept candidates sorted by relevance, where the most relevant concept is at the top of the list. The top-k candidates are taken as the normalization result of the input text. An example of the learning to rank module for the phrase 'Coronary artery disease' can be seen in Figure 6, where the top-1 selection is applied.

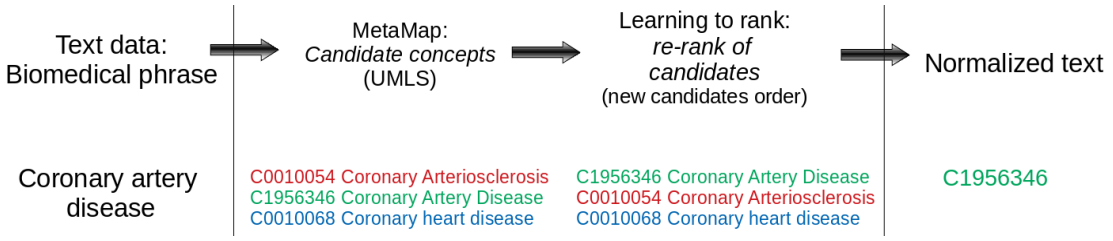

Figure 6. Example of normalization with learning-to-rank.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

The entities that are not assigned any CUI ids then go through our spelling corrector. Our spelling corrector is based on the Levenshtein distance, where it measures the edit distance between the term and the UMLS terms and chooses the most similar UMLS term. Below in Table 2, you can see some examples of the spelling errors and their corrections.

Spelling Error	Correction
insonmia	insomnia
hypercholesterlolemia	hypercholesterolemia
depresion	depression

Table 2. Examples of the L2N spelling corrector.

3.2.3 Results

In this section, we present the performance results achieved by each normalization pipeline. When needed, we also include details concerning the annotated data in which the pipeline is validated.

1. **ZOOMA**

   ZOOMA has been developed to annotate text with ontology terms backed by its curated repository of knowledge. The annotation task of ZOOMA is carried out without any preprocessing. Hence, it expects a preprocessed dataset as an input and otherwise results would not be satisfactory. Here, we did not explicitly conduct any experiments to measure the performance of ZOOMA. Instead, we present the ZOOMA performance in comparison with the SORTA in the next paragraph. The recall is that low because it can only find exact matches, those were often too specific to describe the given input term.

2. **SORTA**

   SORTA creates a ranked list of the best lexical mappings between input terms and Ontology terms based on lexical and semantic mapping. An expert needs to select the correct answer. SORTA was applied to map terms from the Dutch CINEAS coding system to the HPO ontology. (CINEAS is the Dutch national disease code development and distribution center for the clinical genetics community. [9, 11]) The first ranked out of SORTAS multiple suggestions had a precision/recall of 0.58/0.40. For the same dataset & ontology ZOOMA achieved a precision/recall of 0.63/0.17 for it's first suggested term. Another use case for SORTA was the Healthy Obese Project (HOP) where Dutch free text answers about different types of sports needed to be mapped to a Metabolic Equivalent Task (MET). The re-use of manually curated data from a previous coding round resulted in a major increase of SORTAs performance. The recall/precision went from 0.59/0.65 to 0.97/0.98 for the rank 1 suggestion. [9]

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

SORTA has not been applied to the CINECA synthetic cohorts' data since it is an interactive tool that requires an expert to perform the mapping.

3. LexMapr

Mapping representation in LexMapr

The mapping of input descriptions to ontology terms in LexMapr has been represented in terms of three types of matches: "full-term match", "component match" and "no match". These match types have been explained below and are illustrated further in Table 3 with examples taken from the CINECA CHILD cohort dataset.

Full-term match. The full-term match represents a whole or a complete or a total match with the entire chunk of the text of an input phrase mapping with some ontology term. The full-term matches could happen without any processing or application of rules, for example, a sample description "ear pain" matches exactly and without any treatment with an ontology term "ear pain:HP_0030766", as shown in Table 3. It is annotated as a type of "full-term match" and with the rules category as "direct match" (i.e. no rule has been applied). Some matches of full-term type are possible only with some processing or application of some rules. For example, an input description, "fifths disease" maps to the ontology term "erythema infectiosum:DOID_8743" and this full-term match has been achieved by LexMapr with the application of certain rules. Hence, therefore, it is annotated as the type of a "full-term match" and with the rules category "processed match", having rules applied "['Inflection (Plural) Treatment: fifths', 'Synonym Usage']" as has been shown in Table 3.

Component match. This type of match occurs when there is no complete match for the entire chunk of the text of an input phrase with some ontology term, but some component (or components of the phrase) matches with one or more ontology terms. Based on the degree and semantics of the match, the component match could be further described as a "full-term equivalent match" or "partial match". Realistically, there could be cases when all the key constituents of the input phrase match as a component or components with one or more ontology terms, and nothing nontrivial is left without matching. If these kinds of component matches are semantically valid, they are the representation of the full-term matches and could be considered as equivalent to the full-term matches. For example, the input description "dyspepsia and gerd" has its two components mapped to two separate terms 'dyspepsia:HP_0410281' and 'gastroesophageal reflux disease:DOID_8534' (Table 3) and, therefore, has been considered as a "full-term equivalent match". However, these types of matches need to be inferred semantically equivalent to full-term matches for the equivalence to be valid.

The partial matches represent the component matches that could not be considered as "full-term equivalent match" and when not all nontrivial constituents of input

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

phrases map to ontology terms but at least the key constituent of the phrase (normally the underlying noun) gets mapped to some ontology term. For example, in the biosample “skin rash due to teething”, its key constituent “skin rash” maps to the ontology term skin rash:SCDO_0001073 though the other constituent of input phrase “due to teething” remains unmapped (Table 3).

No match. LexMapr considers the resulting match as of a type "No Match" when none of the key constituents of the input phrase maps to some ontology term. For example, the input descriptions in the dataset "discharge" and “spitting up” do not map to any term and thus could be considered of the type "no match".

Specimen description	**Matched ontology terms with ontology ids**	**Match type**	**Match subtype**
ear pain	ear pain:HP_0030766	Full-term match	Direct match
fifths disease	erythema infectiosum:DOID_8743		Processed match Rules: ['Inflection (Plural) Treatment: fifths', 'Synonym Usage']
fever, headaches	'fever:SYMP_0000613', 'headache:HP_0002315'	Component match	Full-term equivalent match
dyspepsia and gerd	'dyspepsia:HP_0410281', 'gastroesophageal reflux disease:DOID_8534'		Full-term equivalent match Rules: [gastroesophageal reflux disease: ['Synonym Usage']
skin rash due to teething	'skin rash:SCDO_0001073'	Component match	partial match
slipped cheek syndrome	'cheek:UBERON_0001567', 'syndrome:DOID_225'		partial match
discharge	-	No match	
spitting up	-		

Table 3. The term mapping types in LexMapr

Evaluation dataset for performance measurement

While the LexMapr’s CINECA-adapted version was trained for rules based on the CHILD cohort synthetic dataset, the evaluation dataset for performance measurement consisted of 600 anonymous and unseen unique descriptions that were obtained from CHILD cohorts’s free-text field describing medication reason. The results from these 600 descriptions were divided among 5 annotators for the evaluation for performance measurement.

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

Evaluation worksheets and annotation guidelines

To facilitate the uniform evaluation, a set of annotation guidelines was prepared. Different evaluation worksheets were created and distributed amongst the annotators. The evaluation sheets along with these guidelines were provided to the annotators for the performance measurement evaluation using a CINECA-adapted version of the LexMapr pipeline. The evaluation has focussed on finding the mapping errors attributed to the LexMapr pipeline only and excluding the errors due to the missing content in ontologies.

Error Characterization: To differentiate the errors due to pipeline functioning from missing content in ontologies or resources, these were classified into two main types - pipeline errors (semantic and functional) and resource errors (Table 4). In the case of pipeline errors, the semantic error (locally referred to as a type A error in the evaluation process) characterizes the match achieved by the pipeline that was considered to be a semantically inaccurate match. The functional error (locally referred to as a type B error) represents the match missed due to pipeline failure or incapability, even when the corresponding ontology term was available. Resource error (locally referred to as a type C error) represents the missed match due to the missing content (terms) in ontologies (Table 4).

This error classification scheme enables the determination of genuine pipeline accuracy. One key aspect in the performance evaluation was to determine whether a component match could be considered as a full-term equivalent match or not. The identification of a correct match (either a full-term match or component match deemed equivalent to a full-term match) helps in distinguishing it from a partial match. This facilitates the calculation of performance metrics based on strict criteria that consider only the full-term match or full-term equivalent match as the correct match and discount the partial matches.

Error type	**Error sub-type**	**Error characterization**	**Error description**	**Input description**	**Example**
Pipeline error	Semantic error (Type A)	Wrong match (semantically/ other reasons)	Reflects that the match is made there by pipeline but it is an inaccurate match.	chile powder	For example, if Chile mapped to a country/ GeoEntity

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

	functional error (Type B)	Missed match to an existing ontology term.	Reflects the match that the pipeline misses even if the term is available in ontologies.	nausea from tonsillitis	Maps nausea to nausea:HP_0002018, but 'tonsillitis' is not matched to tonsillitis:DOID_10456' by the system
Resource error	Term or synonym deficit error (Type C)	Missing ontology content	It represents the missed match due to the content being unavailable in the ontologies.	muscle discomfort	muscle discomfort - there is no term in the selected resources i.e. ontologies

Table 4. Different types of mapping errors used in the LexMapr evaluation.

Evaluation results

The LexMapr performance has been measured by standard measures of recall, precision, and F-measure. Recall, a metric depicting the coverage of mapping achieved, measures here the number of correctly mapped entities as a percentage of the total correct entities. For LexMapr performance evaluation, the coverage has a specific connotation for recall (pipeline recall) that is based only on pipeline's performance for mapping after discounting the missed coverage owing to missing content in ontologies. The aim of this performance evaluation was to calculate the genuine performance of the LexMapr pipeline by not considering the resource errors. The metrics precision and F1-measure have been calculated based on pipeline errors and pipeline recall. It means, for LexMapr performance evaluation, the resource deficit errors (annotated as Type C errors by different annotators) were excluded from the performance calculations.

Therefore, input descriptions considered in calculations after excluding the resource deficit errors were 402 (from a total of 600 test samples, 198 resource errors were excluded). The pipeline errors based on functional errors provide the missed matches due to the pipeline and thus contributing to the calculation of pipeline recall. The semantic errors when taken alone represent the wrong matches used to calculate the pipeline accuracy as these types of errors reflect the wrong or inaccurate mappings done by the pipeline. Table 5 shows the evaluation statistics based on the strict criteria for the test dataset.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

Total samples	Resource errors (missing content)	Samples used for pipeline evaluation	Correct matches	Missing matches due to pipeline functionality	Pipeline recall	Pipeline errors (wrong matches)	Pipeline accuracy	F1-Measure
600	198	402	286	78	78.57	38	88.27	83.14

Table 5. LexMapr pipeline evaluation results based on strict criteria for a dataset describing medication reasons in CHILD cohort dataset.

By looking at the evaluation results using the strict criteria, the pipeline achieved an accuracy of ~88% and F1-measure of ~83% for the CHILD cohort test dataset, apart from providing a recall of ~78%. The high percentage values of metrics suggest that the LexMapr, equipped with proper rules and required lookup resources, succeeded in obtaining the desired mapping in the given domain (provided that the content is available in the selected ontologies). However, if the content missing in the ontologies is included for the calculation of recall and other measures for the system as a whole, these values would certainly be lower. It is worth noting that the missed mapping due to absent ontology content was ~33% (198 out of 600). This highlights that it is very difficult to have complete and up-to-date biological resources (ontologies or other terminological resources). Another important observation was that the spelling of 94 (~16%) descriptions were corrected by BioSpellC out of a total of 600 input phrases. This reflects how messy these short text descriptions are there in the CHILD cohort data.

4. L2N

In this subsection, we present the performance of the L2N pipeline developed at HES-SO/SIB. First, we describe the MCN corpus, containing manual annotation of biomedical concepts, which was used to assess the performance of the L2N pipeline for the concept normalization task. Then, we describe the learning to rank module applied in this pipeline. Finally, we present the normalization results achieved by this pipeline.

Annotated Data

MCN corpus is annotated with UMLS concepts, containing 100 discharge summaries annotated with 3,792 unique concepts including medical problems, treatments, and tests [1]. MCN corpus is divided into 50 discharge summaries for training and 50 discharge summaries for testing. An example of annotation from a passage in MCN is shown in Figure 7. The mention 'coronary artery' is annotated with the CUI C1956346, where the concept name is 'Coronary Artery Disease' in the UMLS metathesaurus.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

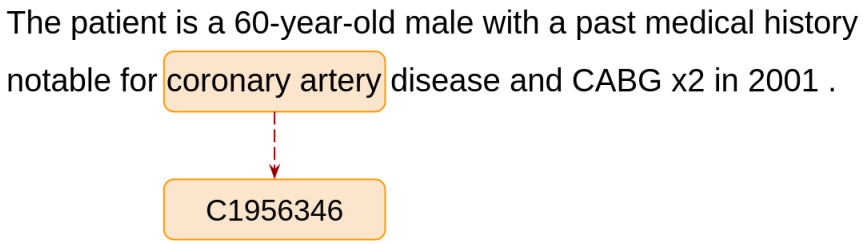

Figure 7. Annotated passage example from MCN. The mention of 'coronary artery' is annotated with the UMLS CUI C1959346.

In UMLS, each concept is associated with a semantic group, e.g., the mention 'High Cholesterol' is annotated with the CUI C0020443 and concept name *Hypercholesterolemia* which belongs to the *Disorders* semantic group. Table 6 shows the distribution of concepts in the UMLS semantic groups in MCN corpus. In this table, #Train and #Test columns represent the number of concepts of the corresponding semantic group while the %Train and %Test columns represent the percentage of concepts in each semantic group with respect to the total of annotated concepts in each set. As we can see, while data has a similar distribution between train and test sets, the distribution of instances across the semantic groups is unbalanced. Some medical mentions are not specific enough to be assigned to a particular concept, e.g., 'no acute distress'; 'multiple medical problems'; 'head , eyes, ears, nose and throat exam'. These mentions in MCN are annotated as CUI-less representing 2% and 3% in the training and test sets, respectively.

Semantic group	#Train	#Test	% Train	% Test
Disorders	2337	2329	34.96%	33.63%
Procedures	1903	1940	28.47%	28.01%
Chemicals & Drugs	947	847	14.17%	12.23%
Concepts & Ideas	825	980	12.34%	14.15%
Anatomy	274	313	4.10%	4.52%
CUI-less	151	217	2.26%	3.13%
Devices	98	108	1.47%	1.56%
Physiology	70	78	1.05%	1.13%
Phenomena	36	36	0.54%	0.52%

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

Living Beings	24	47	0.36%	0.68%
Objects	13	17	0.19%	0.25%
Activities & Behaviors	4	7	0.06%	0.10%
Organizations	2	5	0.03%	0.07%
Occupations	0	1	0.00%	0.01%
Total of concepts	6684	6925	100.00%	100.00%

Table 6. Data distribution organized by semantic group

Learning to rank task

Here, we report the evaluation of our L2N pipeline (described in Section 3.2.2) on the MCN corpus. We use both training and test sets of the MCN corpus combined to perform a stratified 5-fold cross-validation. In each fold, the training set is used to train the learning to rank algorithm, and the test set is used to evaluate the learned model. Since there exist various learning to rank algorithms in the literature, here, we examine different approaches and report their performance [12]. Each approach has a specific loss function. The first approach, pointwise, transforms the problem into a classification problem to predict whether a query is relevant. For instance, the loss function of RankMSE is based on the mean square error. In the pairwise approach, each query is associated with a candidate result, the previous approach does not make this link. In this approach, we explored RankNet and LambdaRank. RankNet adopts cross entropy as loss function and gradient descent as optimization algorithm. LambdaRank is based on RankNet but optimized to learn the ranking score of the candidate for the current query. The last approach is Listwise. We explored ListNet and ListMLE. ListNet is also similar to RankNet but in its loss function is considered the current list of candidates and scores. ListMLE minimizes the likelihood loss function.

The performance of the explored learning to rank algorithms are presented in Table 7 in terms of precision@k, when k is 1, i.e., only the first element of the rank list is considered in this evaluation. RankMSE achieved the best performance in MCN corpus with over 5 points higher than the baseline MetaMap.

Learning approach	Method	Precision@1
Baseline	MetaMap	0.5481
Pointwise	RankMSE	**0.6047**

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

Pairwise	RankNet	0.5698
	LambdaRank	0.5834
Listwise	ListNet	0.5934
	ListMLE	0.5789

Table 7. Performance of the learning-to-rank task over MCN corpus.

Concept normalization task

Table 8 shows the performance of each module of the L2N pipeline described in Section 3.2.2. This pipeline is evaluated in the MCN test set. The dictionary matching module of L2N achieved an accuracy of 0.5187 against the MCN train set while the exact match with UMLS achieved an accuracy of 0.6520. Applying learning to rank to the MetaMap list of candidates, L2N achieved 0.7612 of accuracy.

Pipeline stage	Accuracy
Exact match MCN train set	0.5187
Exact match UMLS	0.6520
MetaMap	0.7421
MetaMap with learning to rank	**0.7612**

Table 8. Performance of concept normalization task over MCN test set.

Based on the predictions of MetaMap with learning to rank (RankMSE), Table 9 organizes the performance according to the UMLS semantic groups. The CoLaus/PsyCoLaus normalization needs are mainly focused on diseases, symptoms, findings, drugs, and procedures. In each UMLS semantic group, several semantic types are defined to get a refined gathering of concepts. Thus, *disorders* semantic group includes diseases, findings, signs, and symptoms, among others. In Table 9 can be seen that the proposed model achieved 78.32% for disorders, 72.27% for procedures, and 87.6% for chemical & drugs semantic groups.

Semantic group	Accuracy
Chemicals & Drugs	0.8760
Concepts & Ideas	0.8673
Activities & Behaviors	0.8571
Living Beings	0.8298

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

Organizations	0.8000
Disorders	0.7832
Procedures	0.7227
Anatomy	0.6070
Devices	0.5926
Objects	0.5294
Physiology	0.5256
CUI-less	0.4055
Phenomena	0.3333
Occupations	0.0000

Table 9. Performance of MCN test set by semantic group

3.2.4 Text mining aggregator API

Each partner within the Text Mining group developed a set of tools independently focused on specific cohorts sharing data within the CINECA collaboration. This reduced the complexity at the development time drastically, but added extra complexity to end users as they have to be familiar with different solutions to use them. To alleviate this, we have developed an aggregator API that exposes the different methods in an integrated environment. An API which can route requests to different tools/pipelines based on the problem and a common output format that is easy to interpret. In the following paragraphs we discuss the design and development process of the API and the interface.

We have followed the API first approach to design and develop the aggregator API. At the start of the design process, inputs and outputs to the system were defined. The unifying API has the ability to annotate either a given text term or a file containing a list of terms. Furthermore, it enables one to choose a model (from a list) with which the term should be annotated and optionally to provide the concept type of the text as a hint to the model. The output of the API contains the text, annotated ontologies, and the confidence of each mapping. The contract was first described using the Open API model. The model was later used to automatically generate server code. Figure 8 depicts a sample input and output of the API of the L2N pipeline developed by HES-SO/SIB.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

```
GET    /annotate?concept=disease&model=HESSO_SIB&text=cancer

{
  "text": "cancer",
  "annotations": [
    {
      "text": "cancer",
      "ontology": {
        "id": "UMLS:C0006826",
        "label": "Malignant Neoplasms"
      },
      "score": 1,
      "source": "HES-SO/SIB"
    }
  ]
}
```

Figure 8. Input and output of the API.

Figure 9 depicts the initial wireframe diagram of the designed user interface. We have envisioned integrating all 4 models developed by the different teams. The concept types will be limited to the concepts supported by each model.

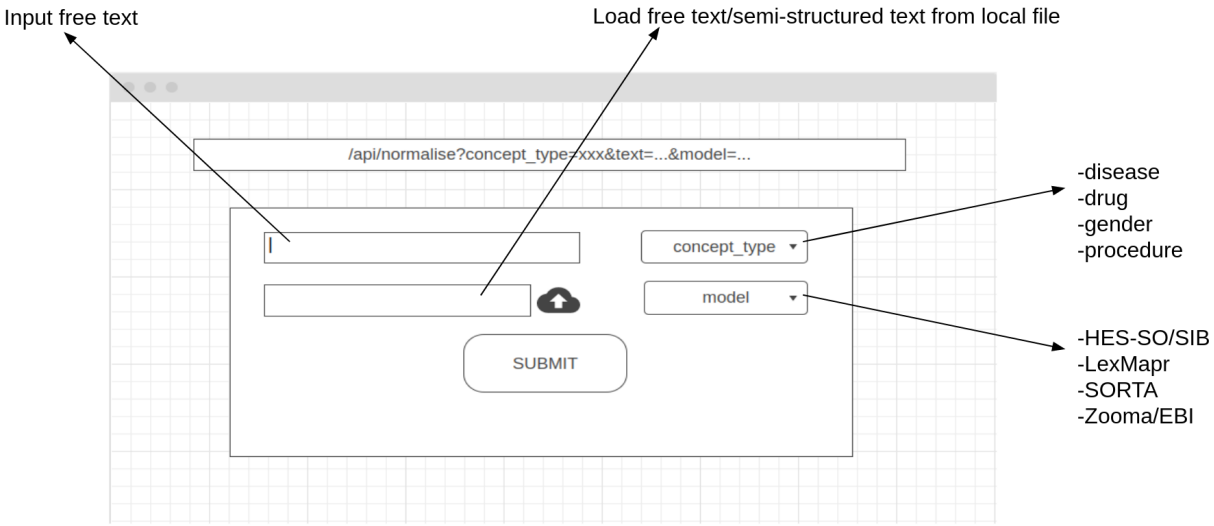

Figure 9. API concept

The software architecture of the aggregator API is depicted in Figure 10. The unified API is defined in openapi.yml[4] file using Open API specification. Server code was generated using the specification file. Based on the selected model the system will route the request to the matching service. Mapping between different services and the unified API was also a part of the aggregator API and was developed by each team.

[4] https://github.com/CINECA-project/wp3-annotator-api/blob/main/spec/src/main/resources/openapi.yml

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

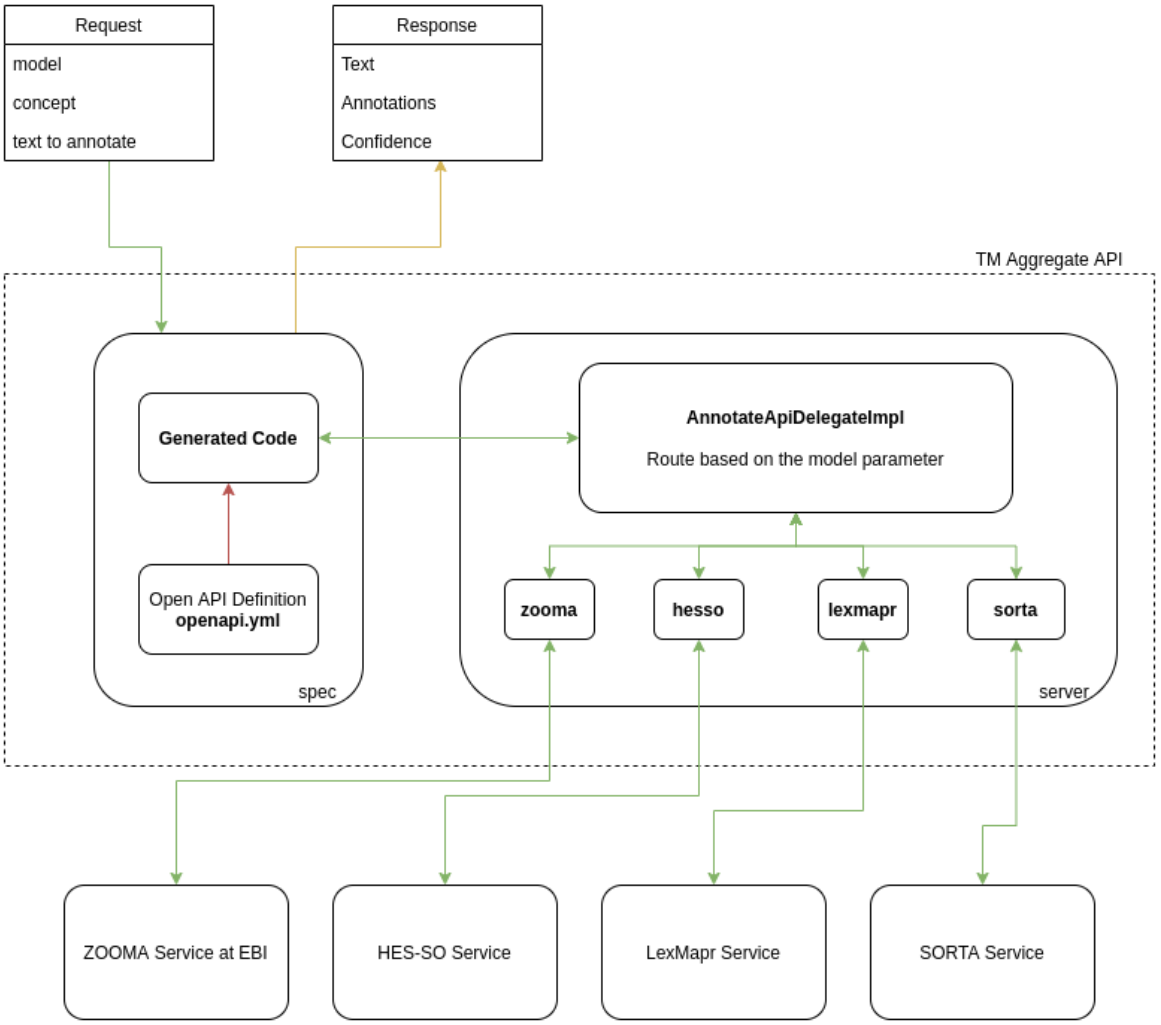

Figure 10. Software architecture of the consolidated API

We have deployed the Aggregator API at EBI Embassy cloud. Currently it is capable of annotating single text terms using ZOOMA and L2N models. We plan to integrate LexMapr and SORTA during the first semester of 2022. Figure 11 and Figure 12 respectively show a sample input and the output of the service.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

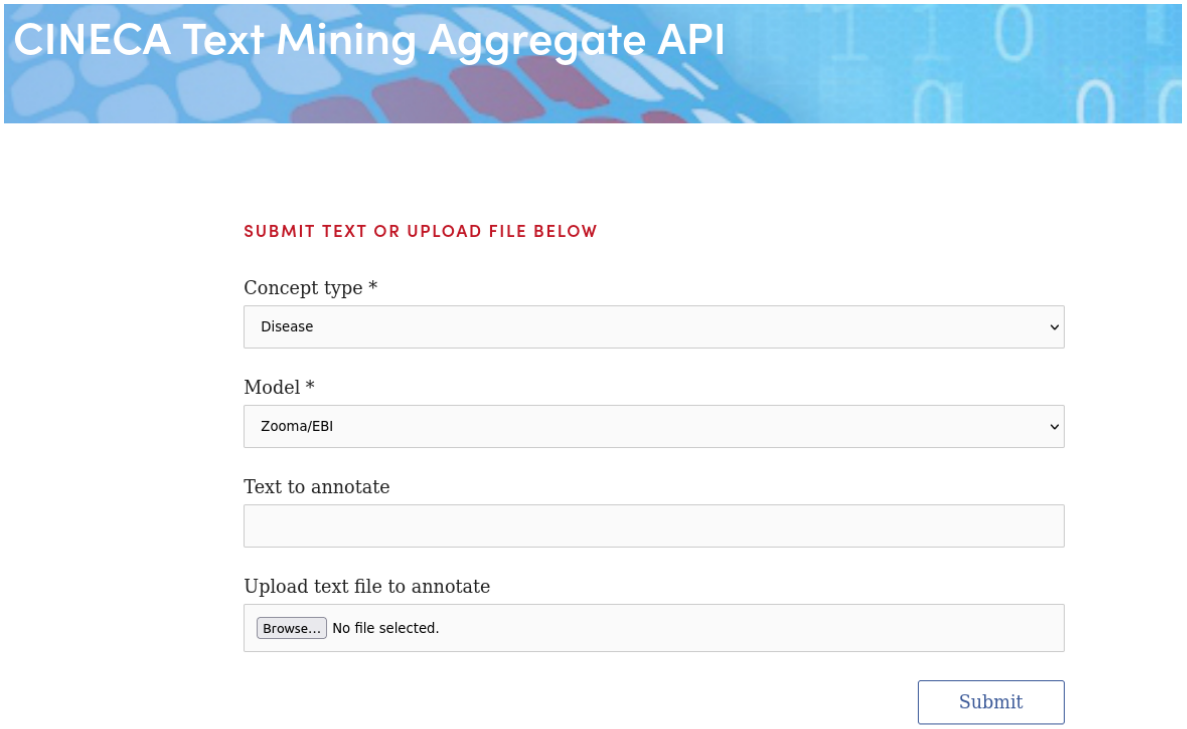

Figure 11. Text mining aggregator API input interface

Figure 12 shows a sample output of the system for the text term "REFLUX, GASTRITIS" annotated by both ZOOMA and L2N models. HES-SO/SIB in the figure refers to the L2N pipeline. The text is normalized to NCI and UMLS terminologies. In NCI the concept name of C57812 is *Gastritis*, CTCAE, while in UMLS the concept name of C4317146 is *Acid reflux* and for C0017152 is *Gastritis*. The confidence of the mapping or a relevance score for each of the results is also attached.

RESULTS

RESULTS FOR THE TERM REFLUX, GASTRITIS

Text	Ontology Id	Ontology Label	Score	Source
REFLUX, GASTRITIS	NCIT:C57812	Gastritis, CTCAE	50	Zooma
REFLUX, GASTRITIS	UMLS:C4317146	Acid reflux	0.3788999915122986	HES-SO/SIB
REFLUX, GASTRITIS	UMLS:C0017152	Gastritis	0.3765999972820282	HES-SO/SIB

Figure 12. Results of the text mining aggregator API for the input text REFLUX, GASTRITIS.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

3.2.5 Concept normalization of free text fields in cohort data

In this section we show the results of our pipelines on CINECA cohort data. After training our models on external datasets, the resulting tools were applied on CHILD and CoLaus/PsyCoLaus synthetic cohorts. Due to the complexity of biomedical entity representation in free-text/semi-structured data, many challenges are faced when normalizing the cohort fields. For example, the free text could contain misspellings, e.g. 'antiinflmmatory', multiple concepts, e.g. 'depression, anxiety and panic disorder', or generic information e.g. 'finished in May 2003'. We started a manual validation of the obtained results to refine our pipelines. Some members of the text mining team are working on medical spelling correctors.

1. **CHILD**

   LexMapr has implemented several functionalities as part of the overall cleaning, normalization and term mapping process for CINECA cohorts free-text fields. Table 10 shows the uses of some of these functionalities in cleaning, normalization and linking free text to ontology terms using LexMapr on medication reason field anonymous values from the CHILD cohort dataset and illustrates the application of different rules to achieve the results.

Input description	**Matched ontology terms with ids**	**Rule applied**
coughs	cough:SYMP_0000614	Singularization
fractured ribs	fractured rib:HP_0041159	
migraines	migraine:DOID_6364	
vaccinatons	vaccination:VO_0000002	Spelling correction treatment
anexity	anxiety:SYMP_0000412	
cronic constipation	chronic constipation:HP_0012450	
hay fever	allergic rhinitis:DOID_4481	Synonym substitution
blocked tear duct	nasolacrimal duct obstruction:HP_0000579	
mild fever	low-grade fever:HP_0011134	

Table 10. Examples showing cleaning, normalization and term mapping results for input phrases from the CHILD cohort dataset

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

2. CoLaus/PsyCoLaus

Free text fields in CoLaus/PsyCoLaus synthetic data are not manually annotated. L2N is applied to these fields, i.e., dginvtx2 and dginvtx3. Some examples of the normalization results are shown in Table 11. The first column corresponds to some examples taken from the synthetic CoLaus/PsyCoLaus free text fields. The next columns show the normalized concept, i.e., UMLS CUI, concept name in UMLS terminology, and the concept semantic group. An example of more than one concept in the input phrase can be seen in Table 11, 'REFLUX, GASTRITIS'. The tool provided two CUIs to map each of the terms included in the input text. Thus, this input text is normalized to C4317146 and C0017152.

CoLaus/PsyCoLaus free text	**UMLS CUI**	**Concept name**	**Semantic group**
OSTEOARTHRITIS	C0029408	Degenerative polyarthritis	Disorders
ARRHYTHMIA	C0003811	Cardiac arrhythmia (Cardiac Arrhythmia)	Disorders
ANTIDIABETIC	C0935929	Antidiabetic agent (Antidiabetics)	Chemicals & Drugs
DENTAL INFECTION	C0877046	Infection of tooth (Tooth Infection)	Disorders
ALLERGY POLLEN	C0018621	Hay fever	Disorders
REFLUX, GASTRITIS	C4317146	Acid reflux	Disorders
	C0017152	Gastritis	Disorders

Table 11. Examples of normalization in CoLaus synthetic data

3.3 Conclusion and next steps

In this deliverable, we presented the methodology used to develop tools to normalize free text in standard medical concepts from CINECA cohorts. Different pipelines were developed independently by partners from EBI, SFU, UMCG, and HES-SO/SIB. Then, they were integrated into a concept annotator API service where the web service exposes individual functionalities of the tools. Some of the pipelines initially were applied to CHILD and CoLaus/PsyCoLaus synthetic data but using the concept annotator API other cohorts will be able to normalize their free text cohort data. Finally, the concept annotator API is a step towards the population of the minimal metadata model.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

An important step towards the population of the minimal metadata model is the alignment with GECKO[5]. In year 4 of the project, we plan to align our outputs with GECKO. We will explore adapting our models to predict using GECKO classes but also use directly ontology mapping tools, such as the Ontology Lookup Service (OLS) developed by EBI. LexMapr and SORTA will be integrated in the concept annotator API. Also, we aim to continue working on each normalization pipeline to get more accurate results in the cohort data which will be available through the main API. The evaluation of spelling corrector solutions also needs further work. Finally, we presented the results each pipeline has achieved on their specific evaluation sets. As a future work, we aim to integrate the evaluation using a common test set.

4. References

[1] Luo, Y.-F., Sun, W., & Rumshisky, A. (2019). MCN: A comprehensive corpus for medical concept normalization. Journal of Biomedical Informatics, 92, 103132. https://doi.org/10.1016/j.jbi.2019.103132.

[2] Aronson, A.R., Lang, F.M.: An overview of metamap: historical perspective and recent advances. Journal of the American Medical Informatics Association17(3),229–236 (2010).

[3] Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M., & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. In Journal of Biomedical Informatics (Vol. 73, pp. 14–29). Elsevier BV. https://doi.org/10.1016/j.jbi.2017.07.012.

[4] Chen, L., Fu, W., Gu, Y., Sun, Z., Li, H., Li, E., Jiang, L., Gao, Y., & Huang, Y. (2020). Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking. Journal of the American Medical Informatics Association, 27(10), 1576–1584. https://doi.org/10.1093/jamia/ocaa155.

[5] Luo, Y.-F., Henry, S., Wang, Y., Shen, F., Uzuner, O., & Rumshisky, A. (2020). The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. Journal of the American Medical Informatics Association, 27(10), 1529-e1. https://doi.org/10.1093/jamia/ocaa106.

[6] Firmann, M., Mayor, V., Vidal, P. M., Bochud, M., Pécoud, A., Hayoz, D., Paccaud, F., Preisig, M., Song, K. S., Yuan, X., Danoff, T. M., Stirnadel, H. A., Waterworth, D., Mooser, V., Waeber, G., & Vollenweider, P. (2008). The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC Cardiovascular Disorders, 8(1). https://doi.org/10.1186/1471-2261-8-6.

[7] Preisig, M., Waeber, G., Vollenweider, P., Bovet, P., Rothen, S., Vandeleur, C., Guex, P., Middleton, L., Waterworth, D., Mooser, V., Tozzi, F., & Muglia, P. (2009). The PsyCoLaus study: methodology and characteristics of the sample of a population-based survey on psychiatric disorders and their association with genetic and cardiovascular risk factors. BMC Psychiatry, 9(1). https://doi.org/10.1186/1471-244x-9-9.

[8] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A.

[5] https://www.ebi.ac.uk/ols/ontologies/gecko

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018.
[9] Pang, Chao; Sollie, Annet; Sijtsma, Anna; Hendriksen, Dennis; Charbon, Bart; Haan, Mark de et al. (2015): SORTA: A system for ontology-based re-coding and technical annotation of biomedical phenotype data. In: Database : the journal of biological databases and curation 2015, S. 1–13. DOI: 10.1093/database/bav089.
[10] Jackson R, Matentzoglu N, Overton JA, Vita R, Balhoff JP, Buttigieg PL, et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database [Internet]. 2021 Oct 26;2021. Available from: http://dx.doi.org/10.1093/database/baab069.
[11] Zwamborn-Hanssen,A.M.N., Bijlsma,J.B., Hennekam,E.F.A.M. et al. (1997) The Dutch uniform multicenter registration system for genetic disorders and malformation syndromes. Am. J. Med. Genet., 70, 444–447.
[12] Li, H. (2011). Learning to Rank for Information Retrieval and Natural Language Processing. Synthesis Lectures on Human Language Technologies, 4(1), 1–113. https://doi.org/10.2200/s00348ed1v01y201104hlt012.

5. Abbreviations

API	Application Programming Interface
cTAKES	Clinical Text Analysis and Knowledge Extraction System
CUI	Concept Unique Identifier
EuCan	European and Canadian consortium projects
GECKO	Genomics Cohorts Knowledge Ontology
MCN	Medical Concept Normalization
NCI	National Cancer Institute
N2C2	National NLP Clinical Challenges
OBO	Open Biological and Biomedical Ontology
OLS	Ontology Lookup Service
OWL	Web Ontology Language
L2N	Learning to Normalize
UMLS	Unified Medical Language System
WP	Work Package

6. Work Packages in CINECA

WP1 - Federated Data Discovery and Querying
WP2 - Interoperable Authentication and Authorisation Infrastructure
WP3 - Cohort Level Meta Data Representation
WP4 - Federated Joint Cohort Analysis
WP5 - Healthcare Interoperability and Clinical Applications
WP6 - Outreach, training and dissemination

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

WP7 - Ethical and legal governance framework for transnational data-sharing
WP8 - Project Management and coordination
WP9 - Ethics requirements

7. Delivery and schedule

The delivery is on time.

8. Appendices

This project has received funding from the European Union's Horizon 2020
Research and Innovation Programme under grant agreement No. 825775.