Knowledge Graph Completion and Enrichment in OntoSides using Text Mining

FINANCÉ PAR ANR
INVESTISSEMENTS D'AVENIR
ANR-16-DUNE-0002

Mohannad Almasri, Fabrice Jouanot, Olivier Palombi (CHU, LJK, UNESS) , Namrata Patel (VISEO), Jose F Rodrigues-Jr, Marie-Christine Rousset

Knowledge Graphs

- Modern knowledge representation formalism based on RDF data model
  - more flexible than the relational model
    - ✓ Extensible schema
    - ✓ No strict separation between schema and instances
  - adapted to data/knowledge sharing between distributed data sources over the Web
    - ✓ the basis of Linked Open Data and the Semantic Web
- Knowledge graph = a set of triples <subject, property, object/value>
  - subject, property and object are URIs (http Uniform Resource Identifiers)
  - **dereferencable URIs** (pointers to Web pages) versus **local URIs**
  - value is a literal (string, integer, date, boolean)

Predefined standard properties

- rdfs:seeAlso
  - relates an URI to the URL of a web page
- rdf:label
  - associates a textual label to an URI
  - a good practice is to associate a label to every URI (possible to associate one label per natural language)
- rdfs:comment
  - associates a short text as a comment for an URI

=> structured data combined with textual content in a u

model

Example: RDF modeling **multiple choice questions** in OntoSides

Q30986 has_for_textual_content **"Concernant la péritonite appendiculaire, donnez la ou les propositions exactes :"** ;

is_linked_to_the_medical_speciality **digestive_surgery**

has_for_proposal_of_answer **prop98552 [** has_for_textual_content **"les signes infectieux sont présents d'emblée » ;**

has_for_correction **« true »]**

prop98553 [has_for_textual_content **"il n'y a pas de défense abdominale ou de contracture" ;**

has_for_correction **« false »]**

prop98604[has_for_textual_content **"elle peut se présenter comme une occlusion fébrile" ;**

has_for_correction **« true»]**

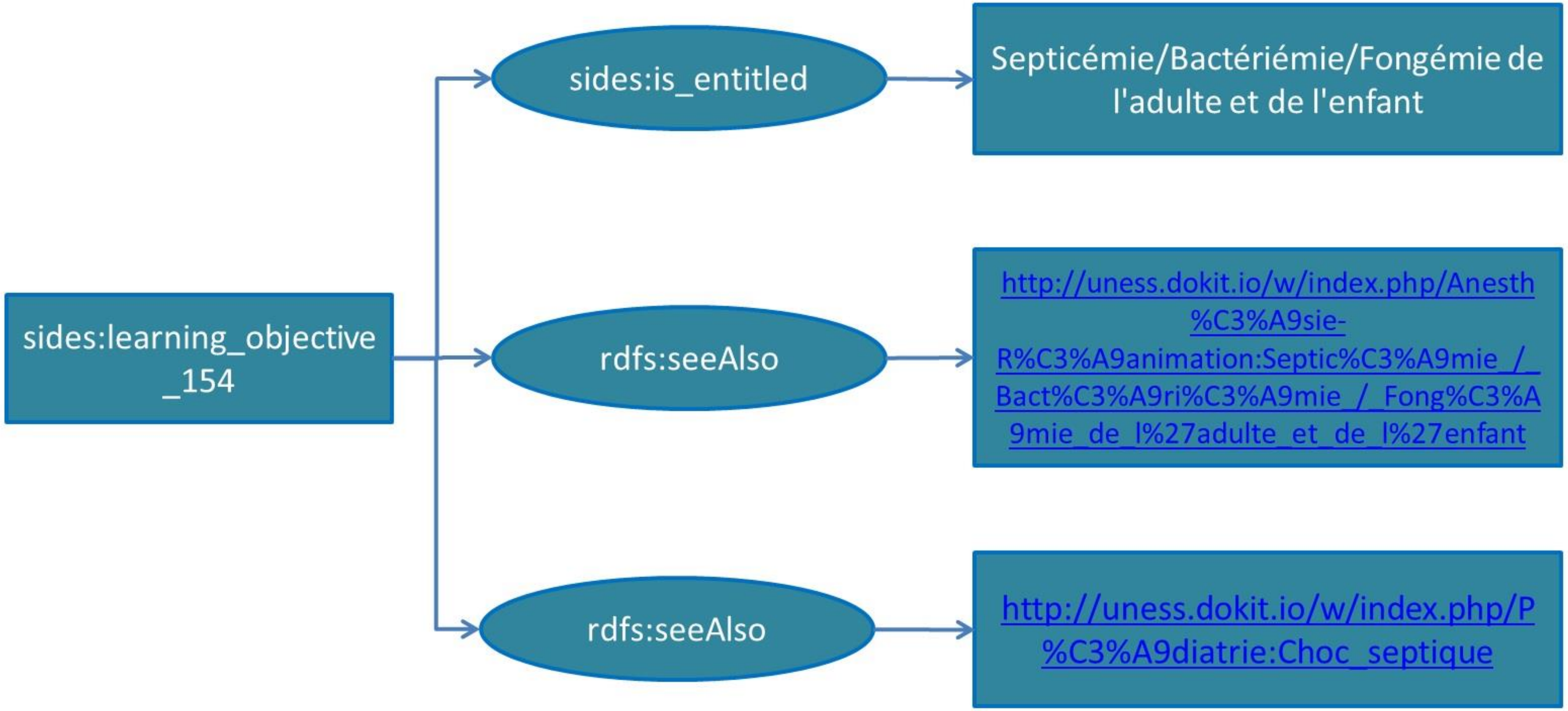
prop98605[has_for_textual_content "**il n'y a pas de pneumopéritoine" ;**

has_for_correction **« true»]**

prop98606[has_for_textual_content **« le traitement est chirurgical" ;**

has_for_correction **« true»]**

Example: modeling learning objectives with seeAlso links to web pages of the wiki SIDES (*)



(*) official educational content provided by the association of French Medical colleges covering the F
program examination

Knowledge graph construction

- Automatic data extraction
  - from text : **DBpedia, Yago**
  - from existing databases: **OntoSIDES**
    - using **mappings** from a **source database schema** to a **target ontology**

⇒**may be very large but still incomplete by nature**

  - properties partially filled

⇒**Knowledge graph completion** has become a problem of increasing interest for which several **supervised** and **unsupervised** techniques have been investigated

OntoSIDES knowledge graph

- the data and knowledge layer of the SIDES 3.0 learning management system for medical studies in France
  - describes **training and assessments activities** performed by more **than 145,000 students** in Medicine **over almost 6 years**
    - exams and training tests are made of multiple choices questions
    - students activities are described at the granularity of **time-stamped clicks of answers** done by students for choosing among the proposals of answers (correct or distractors) associated to questions

⇒**6,5 billions triples** with almost **400 millions clicks** coming from the answers of students to almost **1,4 million questions**.

  - **13% questions** have been explicitly **linked** by their authors to **medical specialties**
  - **12% questions linked** to **learning objectives** (items listed in the Fren
    medical reference program)

Data incompleteness

- Problematic for conducting **well-grounded learning analytics**
  - partial answers for basic Select From Where queries
  - **wrong results for aggregate or counting queries**

- This may occur on some specific properties likely to be involved in aggregate queries to define dimensions
  - is_linked_to_medical_specialty (from questions to medical specialties)
  - is_linked_to_ECN_item (from questions to learning objectives)

Knowledge graph completion and enrichment

- Knowledge graph completion
  - automatically inferring missing facts from existing ones
    - between **questions** and **medical specialties** or **learning objectives**
- Knowledge graph enrichment
  - Automatically discovering links with external reference knowledge graphs or standard ontologies
    - Standard **UMLS (Unified Medical Language System)** medical terminologies like **MeSH (Medical Subject Headings)** and **SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms)**

=> Can be modeled as **classification** or **matching** problems

- depending on the available textual description of the target entities and availability of training data

Automatic discovery of missing links in OntoSIDES

Inferring links between questions and medical specialties

- can be solved as a **multi-label** and **multi-class classification** problem
  - **multi-class**: 31 possible classes (the different medical specialties)
  - **multi-label**: question can be linked to more than one medical specialty
  - **training set**: the 149,000 (13%) questions (with their textual description) for which the property is linked to the medical specialty is valued
- using several classifiers
  - Naive Bayes, Maximum Entropy, CNN (Convolutional Neural Network)

Automatic discovery of missing links in OntoSIDES

Inferring links between questions and learning objectives

- a **multi-label** and **multi-class classification** problem
  - **multi-class:** 362 possible classes (the different learning objectives) and **multi-label**
  - **training set** : 144,000 (12%) questions for which the corresponding property is valued
- can be also solved as a **matching** problem between the textual descriptions of the questions and of the learning objectives
  - only for the **236 learning objectives** that have a textual description
- using several variants of TF-IDF ranking function used in **Information Retrieval** to return the top-k learning objectives for each

Comparative experimental results for classification

Dataset	Classifier	Hits@1	Hits@2	Hits@5	Hits@10	MRR
Dataset1	Naive Bayes classifier	73.8%	83.1%	84.2%	84.3%	79.9%
	Maximum Entropy classifier	75.1%	88.9%	95.4%	96.8%	84%
	CNN classifier	76.4%	89.4%	96.3%	98.5%	85.2%
Dataset2	Naive Bayes classifier	56.4%	64.8%	67.8%	67.9%	61.5%
	Maximum Entropy classifier	68%	81.7%	90.6%	93.6%	78.2%
	CNN classifier	66.4%	78.9%	88.8%	93.4%	76%

Dataset1: 149145 questions -> 31 medical specialties
Dataset2: 144708 questions -> 362 learning objectives
Hits@k (Precision at k): average number of times a correct result appears within the top-k answers
MRR (Mean Reciprocal Rank): average of the rank inverses of the first correct answer

- **All the classifiers perform better on Dataset1 than on Dataset2**
  - the number of classes for Dataset2 is more than 10 times the number of classes for Dataset1 for almost the same number of items to classify
- **Naive Bayes outperformed by Maximum Entropy and CNN**
- Maximum Entropy gives slightly better results than CNN classifier on Dataset2
  - **in more than 96% (93%) of the cases, the correct medical specialties (learning objectiv**
    the top-10 answers

Application: proposing suggestions to teachers while editing a new question

Édition d'une nouvelle question

Intitulé de la question

Dans la situation d'une pyélonéphrite aiguë de l'enfant avec présence de cocci gram positif en chainettes à l'examen direct d'un ECBU, dire parmi les traitements, lequel (lesquels) est (sont) initialement recommandé(s) :

Réponses

cotrimoxazole par voie orale	Faux
céphalosporine de 3° génération (ceftriaxone) p	Faux
pénicilline A (amoxicilline) par voie parentérale	Valide
pénicilline M (oxacilline) par voie parentérale	Faux
pénicilline A (amoxicilline) par voie orale	Faux

Spécialité(s)

Propositions du système:

- Pédiatrie
- Maladies infectieuses

Ajoutez une spécialité:

- Addictologie
- Anesthésiologie - Réanimation - Urgences
- Cancérologie - Radiothérapie
- Cardio-vasculaire
- Dermatologie
- Chirurgie digestive
- Endocrinologie - Métabolisme - Nutrition
- Médecine légale
- Gérontologie

>

<

<<

Spécialité(s) que vous retenez:

- Pédiatrie

Comparative results of classification and matching

Dataset3: 108818 questions -> 236 learning objectives with textual description

Method		Hits@1	Hits@2	Hits@5	Hits@10	MRR
unsupervised	Jelinek-Mercer applied to bags of words	44.5%	58.2%	72.2%	80.9%	57%
	BM25 applied to bags of semantic terms	51.5%	64%	75.7%	81.9%	62.4%
supervised	Naive Bayes classifier	56.2%	64.6%	67.5%	67.7%	61.3%
	Maximum Entropy classifier	68.2%	81.4%	90.4%	93.6%	78%
	CNN classifier	66.2%	78.9%	88.6%	93%	75.9%

- **The "bag of semantic terms" representation** leads to more accurate results than **the "bag of words" representation**
  - **Semantic terms** are **medical concepts** that are **automatically extracted** from the textual descriptions **by** using the **SIFR BioPortal Annotator** (LIRMM, Clément Jonquet) applied to the French versions of the reference biomedical terminologies MESH and SNOMED
- Not surprisingly, **supervised classification methods outperform the unsupervised ones** (except Naïve Bayes at precision 5 and 10)
- However, **unsupervised methods provide good results (above 80%) at precision 10**

Automatic Discovery of Links with External Ontologies

- From the 236 **learning objectives** with a textual description to standard **medical concepts** described in biomedical ontologies
  - UMLS concepts in MeSH and SNOMED CT (French and English version)
- Method overview
  - applied to the set of learning objectives, **seen as a corpus**, each learning objective being seen as a document described by a bag of medical concepts
    - Computation of the term frequency (TF) and the inverse document frequency (IDF) for each medical concept present in the corpus
    - Filtering out the medical concepts with a low IDF (below a certain threshold fixed experimentally)
    - For each learning objective, return the top-k medical concepts (ordered by decre
      k is also fixed experimentally

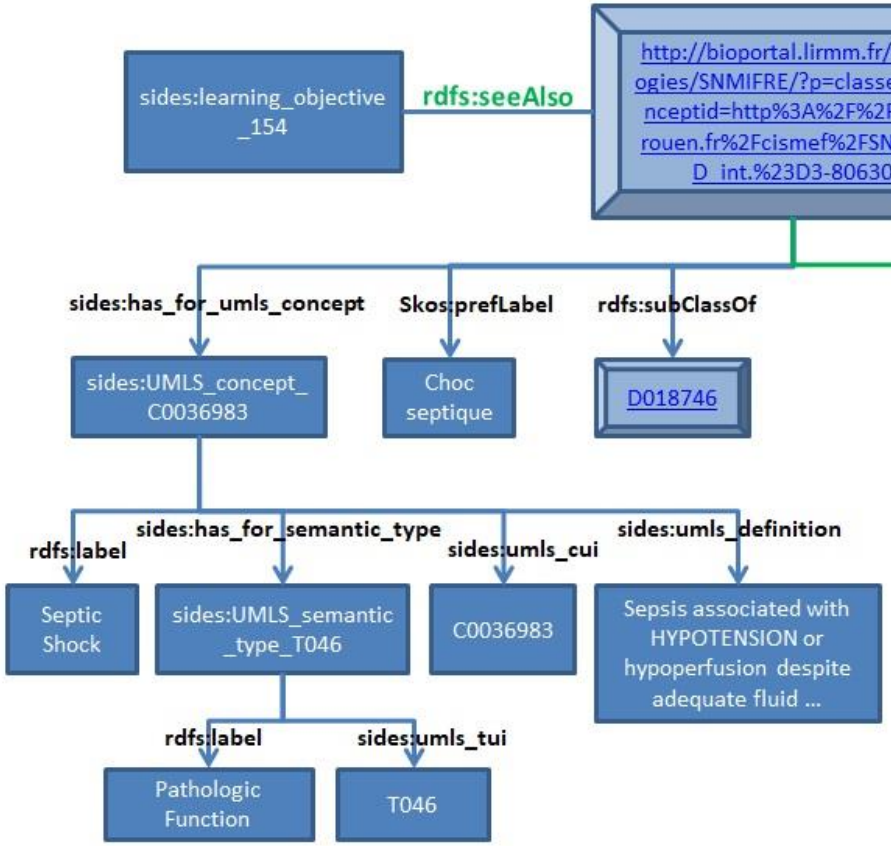
Two-step validation: method and results

- No training dataset available
- First step of validation on 15 learning objectives with a domain expert (O.Palombi)
  - calibration of the parameters to get the best precision at precision k
- Second step of validation with medical experts through an online validation interface
  - answers of experts on 96 learning objectives

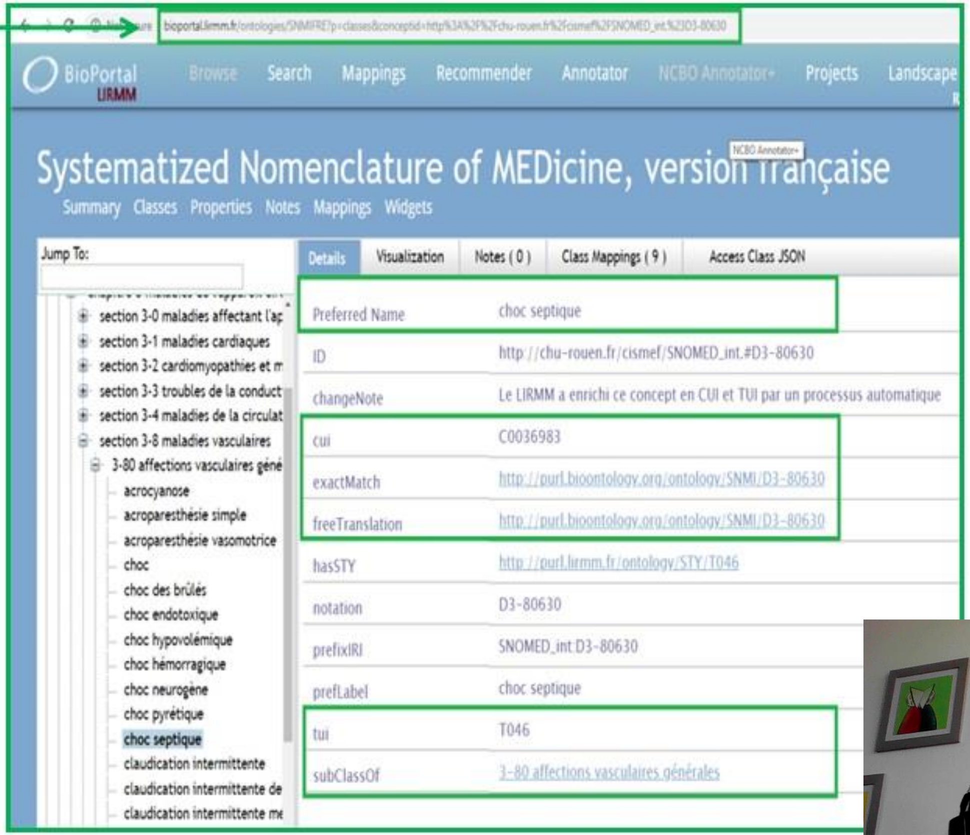
#Evaluated Learning Objectives	#MSHFRE and SNMIFRE Semantic Terms	P@5
96	510	94.5%

  - OntoSIDES enrichment by adding useful triples in addition to the disco links
    - 15371 triples added

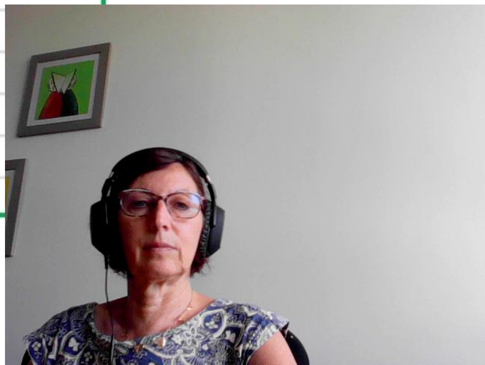
Example



sides:learning_objective _154
rdfs:seeAlso
http://bioportal.lirmm.fr/ontologies/SNMIFRE/?p=classes&conceptid=http%3A%2F%2Fchu-rouen.fr%2Fcismef%2FSNOMED_int.%23D3-80630
sides:has_for_umls_concept
Skos:prefLabel
rdfs:subClassOf
sides:UMLS_concept_ C0036983
Choc septique
D018746
rdfs:label
sides:has_for_semantic_type
sides:umls_cui
sides:umls_definition
Septic Shock
sides:UMLS_semantic _type_T046
C0036983
Sepsis associated with HYPOTENSION or hypoperfusion despite adequate fluid ...
rdfs:label
sides:umls_tui
Pathologic Function
T046



BioPortal
LIRMM
Search
Mappings
Recommender
Annotator
Projects
Landscape
Systematized Nomenclature of MEDicine, version française
Summary Classes Properties Notes Mappings Widgets
Jump To:
Details
Visualization
Notes (0)
Class Mappings (9)
Access Class JSON
section 3-0 maladies affectant l'ap
section 3-1 maladies cardiaques
section 3-2 cardiomyopathies et m
section 3-3 troubles de la conduct
section 3-4 maladies de la circulat
section 3-8 maladies vasculaires
3-80 affections vasculaires géné
acrocyanose
acroparesthésie simple
acroparesthésie vasomotrice
choc
choc des brûlés
choc endotoxique
choc hypovolémique
choc hémorragique
choc neurogène
choc pyrétique
choc septique
claudication intermittente
claudication intermittente de
claudication intermittente me
Preferred Name
choc septique
ID
http://chu-rouen.fr/cismef/SNOMED_int.#D3-80630
changeNote
Le LIRMM a enrichi ce concept en CUI et TUI par un processus automatique
cui
C0036983
exactMatch
http://purl.bioontology.org/ontology/SNMI/D3-80630
freeTranslation
http://purl.bioontology.org/ontology/SNMI/D3-80630
hasSTY
http://purl.lirmm.fr/ontology/STY/T046
notation
D3-80630
prefixIRI
SNOMED_int:D3-80630
prefLabel
choc septique
tui
T046
subClassOf
3-80 affections vasculaires générales



Conclusion

- Specific completion and enrichment problems
  - targeting property of interest guided by the needs in data analytics of domain experts.
- Generic methodology
  - exploiting textual information found in knowledge graphs through datatype properties or rdfs:seeAlso links to web pages.
- Experimental results
  - demonstrated that it can effectively perform big knowledge graph completion and enrichment with a precision up to 95%