NoNLP: Annotating Medical Domain by combining NLP techniques with Semantic Technologies

Ghislain Auguste Atemezing[1]

Mondeca, 35 Boulevard Strasbourg, 75010, Paris, France
ghislain.atemezing@mondeca.com

Abstract. We present in this work the process of annotating data from the medical domain using gazetteers used as reference for the annotation. The process combines semantic web technology and NLP concepts. The application is proposed in this eHealth challenge for multilingual extraction of IC10 codes. The first results give some directions on which aspects of the workflow to improve to make a better system.

Keywords: eHealth, RDF, NLP, ICD-10, multilingual extraction, semantic annotation, death certificates

1 Introduction

This working note presents the approach used to annotate and detect WHO ICD-10 codes in two datasets of death certificates: one in French and another one in English. The work has been done during the eHealth challenge 2017 in the task 1 for multilingual information extraction task [7]. eHealth challenge is part of the labs in the CLEF 2017 [4] conference in the fields of multilingual and multimodal information access evaluation. The present document describes the tasks performed in Section 2, followed by the objectives to achieve by the experiments in Section 3. Section 4 provides the description of the approach Section 5 gives an overview of the resources used by our approach. Furthermore, we provide the results in Section 6, with a brief analysis of some insights in Section 7 before giving some perspectives to improve the current work.

2 Tasks Performed

We have performed the following tasks:

- Transform all the datasets into the RDF [2] model for a graph-based manipulation
- Transform/convert the dictionaries received for the challenge for all the years into the Simple Knowledge Organisation vocabulary SKOS [6] for better enrichment across the knowledge-base.

– Design a GATE [3, 5] workflow to annotate the RDF datasets based on Gazetteers extracted from the dictionaries
– Work on both French (raw data) and English corpus on a single workflow, in a multilingual approach. Thus, we are able to handle more languages.

3 Main objectives of experiments

The main objectives are to retrieve the relevant ICD-10 codes in a text field of a CertDC document line by line as they are provided in the dataset.

4 Description of the Approach

The so-called approach "Not Only NLP" (NoNLP) is a combination of NLP technique for entity extraction from text based on GATE annotator and the extensive use of RDF model for data manipulation within the system and to further enrich the data as a knowledge base. NLP helps using GATE pipeline We use the Content Augmentation Manager (CA-Manager) [1], a semantic tool for knowledge extraction from unstructured data (text, image, etc) and a knowledge manager, able to handle both the model and the data in RDF.

4.1 Content Annotation

The annotation process employed is based on a central component: the Content Augmentation Manager (CA-Manager). CA-Manager is in charge of processing any type of content (plain text, XML, HTML, PDF, etc). This module extracts the concepts and entities detected using text mining techniques with the text input module. The strength of CA-Manager is to combine semantic technologies with a UIMA-based infrastructure[1] which has been enriched and customized to address the specific needs of both semantic annotation and ontology population tasks.

In the scenario presented in this paper, we use the GATE framework for the entity extraction. CA-Manager uses an ontology-based annotation schema to transform heterogeneous content (text, image, video, etc.) into semantically-driven and organized one. The overall architecture of CA-Manager is depicted in Figure 1. We first create the gazetteer with the SKOS document obtained from the experts. We then launch in parallel 10 documents in multi-threads containing text information represented by each row in the CSV. The annotation report contains the valid knowledge section, an RDF/XML document containing the Uniform Resources Identifiers (URIs) of the concepts detected by the annotator. Finally, a SPARQL update query is launched to update the dataset containing the all the data in RDF.

[1] Unstructured Information Management Architecture (`http://uima.apache.org`)

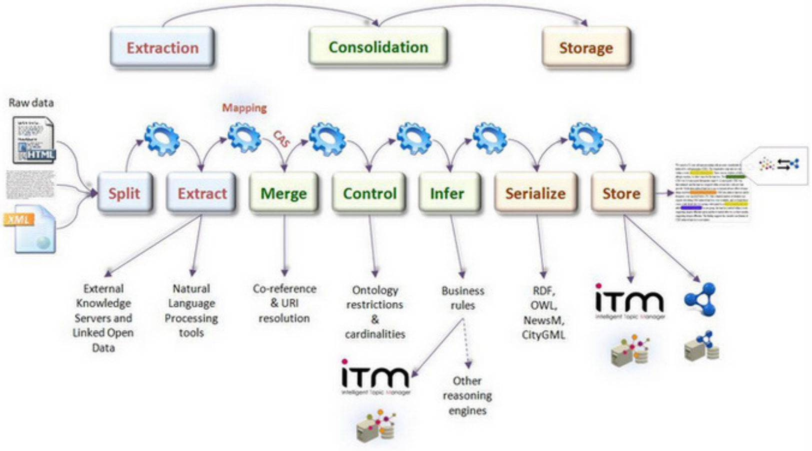

Extraction
Consolidation
Storage
Raw data
HTML
XML
Mapping
CAS
Split
Extract
Merge
Control
Infer
Serialize
Store
External Knowledge Servers and Linked Open Data
Natural Language Processing tools
Co-reference & URI resolution
Ontology restrictions & cardinalities
Business rules
RDF, OWL, NewsM, CityGML
ITM Intelligent Topic Manager
Other reasoning engines

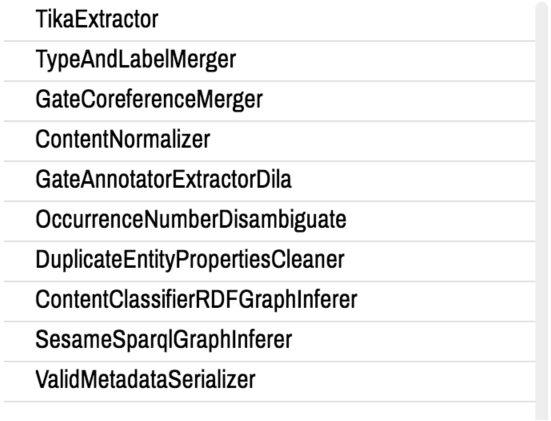
Fig. 1. Pipeline of annotation using CA-Manager

4.2 CA-Manager Configuration

We use an approach composed of three main steps:

1. First we generate the gazetteer to be used for the annotations by SKOS-ification process.
2. Second: we prepare a workflow suitable for our use case, such as defining how to deal with languages, text accentuation, etc. (Figure 2). This includes:
   - the conversion of all the test data from CSV into RDF by using a model (ontology) to transform them into RDF.
   - The actual annotation to extract the pertinent concepts to enrich the initial dataset of test data into an endpoint.
3. Third: We proceed then to enriching the dataset in the endpoint by associating each URI detected in the annotation phase with its relevant code number (which is a property) of the SKOS concept. Finally, a SPARQL CONSTRUCT query is then launched to create the result output in CSV based on the specification of the challenge.

Figure 3 provides a big picture of the NoNLP approach. The process starts with the data conversion in RDF of the dictionaries to generate the gazetteers using SKOS2GATE. The dictionaries is used as configured in the GATE workflow for the annotation by CA-Manager, together with the test dataset already converted into RDF. The result is a set of files representing the valid knowledge in RDF/XML with the entities detected and the associated URI as described in the Gazetteer dataset. All the files output of CA-Manager are merged into the triple store where each piece of information is manipulated by an URI with the associated ICD-10 code if available. The enriched dataset can be exported into CSV according to the specifications of the challenge



TikaExtractor

TypeAndLabelMerger

GateCoreferenceMerger

ContentNormalizer

GateAnnotatorExtractorDila

OccurrenceNumberDisambiguate

DuplicateEntityPropertiesCleaner

ContentClassifierRDFGraphInferer

SesameSparqlGraphInferer

ValidMetadataSerializer

Fig. 2. CA-Manager Workflow configuration used during this challenge

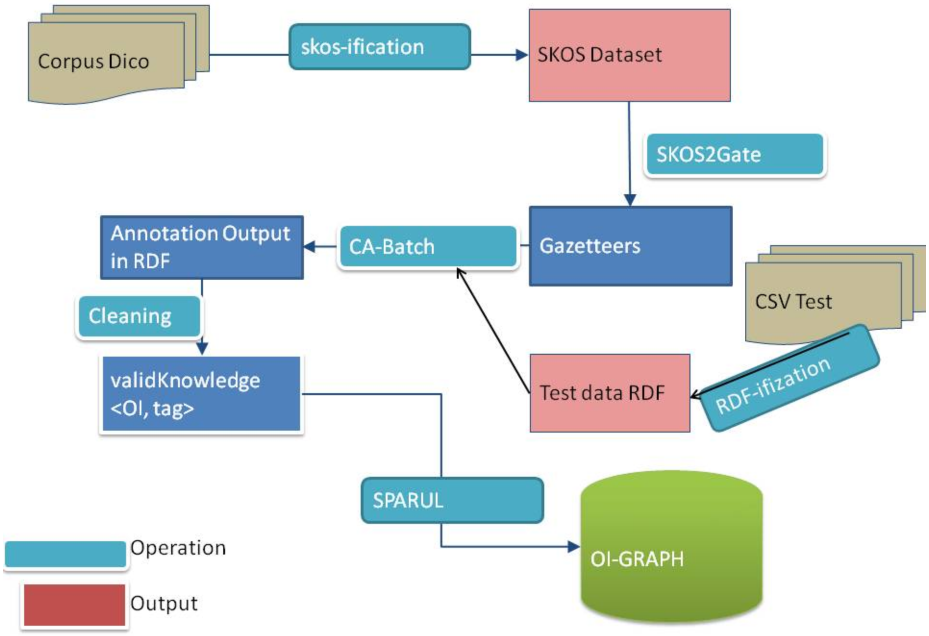


Fig. 3. Overview of the NoNLP approach containing the main operations and the outputs at each step. The architecture combines both NLP techniques and Semantic technologies to detect relevant concepts in pieces of text.

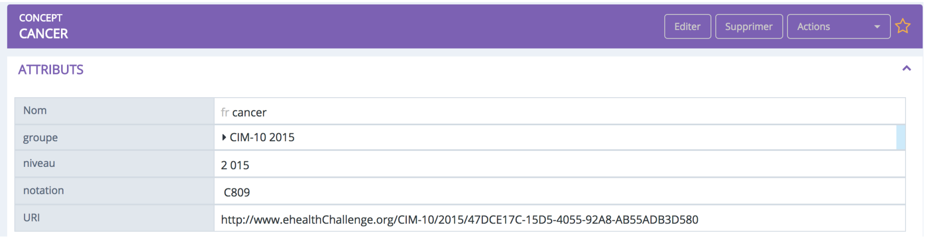
4.3 Gazetteer Generation

We first model all the dictionaries received in RDF using SKOS vocabulary[2]. Each element in the dictionary is a `SKOS:Concept`, and different concepts in

[2] `https://www.w3.org/TR/skos-reference/`

different years have different `skos:inScheme` property. Figure 4 shows a sample view for the concept CANCER modeled in SKOS with the attributes attached to describe the concept in our scenario. This dataset is used as input of our tool SKOS2GATE which transforms the RDF file into list of terms or Gazetteers with some normalization process within the configuration. The configuration file used for creating the dictionary contains two major information:

- Name of the Java class of the morphological analyzer used to lemmatize labels, one per language.
- The indication to convert all the labels to lower case in the dictionary. This is the only normalization we made that affect the dictionary. The choice is guided by the characteristics of the text being analyzed, as we are using exact matching with the terms during the entity detection process.



Nom	fr cancer
groupe	▸ CIM-10 2015
niveau	2 015
notation	C809
URI	http://www.ehealthChallenge.org/CIM-10/2015/47DCE17C-15D5-4055-92A8-AB55ADB3D580

Fig. 4. View of the concept CANCER in SKOS, present in the dictionary in 2015.

4.4 Data Modeling

NoNLP assumes that all the data manipulated are graphs. So, we transform into RDF all the test data to be used in our experiments. The benefits here is that each element of the graph is described by a unique URI to identify a single resource and to be used for merging information attached to it. The input of the annotator is not anymore a document, but a concept with properties representing the actual document to be processed.

5 Resources

We solely used the raw data received without additional extra data. Hence we used the following datasets in their original version received for the challenge:

- All the dictionaries in CSV for French dataset from 2006 to 2015.
- The test dataset “AlignedCauses_2014test.csv” in the French dataset.
- The American dictionary provided in CSV for the English terms.
- The test corpus with data for 2015

	Precision	Recall	F-measure
NoNLP-run	0.691	0.309	0.427
Average	0.670	0.582	0.622
Median	0.646	0.606	0.625

Table 1. Results of the run for English raw corpus

	Precision	Recall	F-measure
NoNLP-run	0.3751	0.1305	0.1936
Average	0.4747	0.3583	0.4059
Median	0.5411	0.4136	0.508

Table 2. Results of the run for French raw corpus

However, we do not use the training data, which is one of the weakness of the approach as described in the working notes. We left that for future work as it will help detect patterns to write suitable JAPE rules for the annotator.

6 Results

With the help of the organizers, the scores of the output of our runs were computed using the evaluation program. The unofficial scores were obtained after converting them to the expected challenge csv format. The scores are as follows:

6.1 EN-RAW Data

We ran the EN corpus containing 14,833 pieces of text in our annotator. The goal was to find in each text the ICD-10 SKOS concept present in the Gazetteer by using an `exactMatch` approach. The score for this dataset is presented in Table 1. We obtained a precision of 69% and a recall of almost 31%.

6.2 FR-RAW data

We ran the FR corpus (raw dataset) containing 59,176 pieces of text in our annotator. The results in Table 2 show a low precision (37.51%) compared to the EN run.

7 Analysis of the results

The results show that we detect better for the English corpus than for the French one. We obtained both high precision and recall with the English corpus, while in French corpus we obtained lower recall. Additionally, our approach seems to work better in English corpus than French, by a factor of 2x. This shows that our system can benefits from the training dataset by adding more alternative labels in general, and in particular by adding some extra normalization based on some patterns that could be observed in the training dataset. We can improve the current scores if we use the training dataset received during this challenge. Our current approach did not make use of the development data so as to help detect patterns to use for creating pattern-matching grammars (JAPE) [8] rules. This can be seen as the basic result that just needs further tuning when exploiting training dataset. We need to understand better the French dataset to enrich our Gazetteers.

8 Future Work

The approach presented in this working notes does not make use of all the power of Semantic technologies, such as the use of the classification rules or inference. We have considered the resources without any hierarchy or relations such as broader, narrower, etc. It could have been possible to have more detected entities based on the relationships within the IC10 concepts. Also, we do not use the training set to add a "learning module" both to improve the gazetteers and then to detect more ICD-10 code. We plan to add JAPE rules based on the patterns detected in the "Gold standard" to improve the GATE workflow detection, and complete the normalization of the gazetteers with additional variations in the label.

References

1. H. Cherfi, M. Coste, and F. Amardeilh. Ca-manager: a middleware for mutual enrichment between information extraction systems and knowledge repositories. In *4th workshop SOS-DLWD "Des Sources Ouvertes au Web de Données*, pages 15–28, 2013.
2. W. W. W. Consortium et al. Rdf 1.1 concepts and abstract syntax. 2014.
3. H. Cunningham. Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254.
4. L. Goeuriot, L. Kelly, H. Suominen, et al. Clef 2017 ehealth evaluation lab overview. In *Lecture Notes in Computer Science*. Springer, Springer, Berlin / Heidelberg, German (in press), 2017.
5. T. Kenter and D. Maynard. Using gate as an annotation tool. *University of Sheffield, Natural language processing group*, 2005.
6. A. Miles, B. Matthews, M. Wilson, and D. Brickley. Skos core: simple knowledge organisation for the web. In *International Conference on Dublin Core and Metadata Applications*, pages 3–10, 2005.
7. A. Névéol, R. N. Anderson, K. B. Cohen, C. Grouin, T. Lavergne, G. Rey, A. Robert, C. Rondet, and P. Zweigenbaum. Clef ehealth 2017 multilingual information extraction task overview: Icd10 coding of death certificates in english and french. In *CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*. Springer, Springer, Berlin / Heidelberg, German (in press), September 2017.
8. D. Thakker, T. Osman, and P. Lakin. Gate jape grammar tutorial. *Nottingham Trent University, UK, Phil Lakin, UK, Version*, 1, 2009.