Enrichment/Population of customized CPR (Computer-based Patient Record) ontology from free-text reports for CSI (Computer Semantic Interoperability)

Mendes, David*; Rodrigues, Irene Pimenta*; Solano-Rodriguez**, Carlos; Baeta, Carlos Fernandes†

* Universidade de Évora; ** Universidad de Alcalá de Henares; † Hospital José Maria Grande

Abstract. CSI (Computer Semantic Interoperability) is a very important issue in healthcare. Ways for heterogeneous computer systems to “understand” important facts from the clinical process for clinical decision support are now beginning to be addressed. We present here comprehensive contributions to achieve CSI. EHR (Electronic Health Record) systems provide a way to extract reports of the clinicians activity. In order to formalize an automated acquisition from semi-structured, free-form, natural language texts in Portuguese into a Clinical Practice Ontology an important step is to develop the ability of decoding all the nicknames, acronyms and short-hand forms that each clinician tend to write down in their reports. We present the steps to develop clinical vocabularies extracting directly from clinical reports in Portuguese available in the SAM (Sistema de Apoio ao Médico) system. The presented techniques are easily further developed for any other natural language or knowledge representation framework with due adaptations.

1 Work in progress and motivation

We will present in this paper two sections that illustrate our work:

1. The steps that we consider are involved in the complex acquisition procedure of clinical concepts expressed in English from text in Portuguese.
2. A proposal of a software architecture needed for an automated acquisition to articulate the steps presented before.

Our work final objective is to enrich/populate an ontology that shall allow us to devise AI (Artificial Intelligence) tools that reason about clinical practice. Given the reasons explained in [4] we chose CPR (Computer-based Patient Record) Ontology as target for population. CPR is a W3C (World Wide Web Consortium) standard for representing clinical practice knowledge. The major problem of personal jargon creation was not properly addressed, however, and so is fully discussed in the present article. We demonstrate the possibility of information extraction from free-text clinical episode reports in an automated manner.
When developing a methodology for automatic population of a CPR Ontology we faced the particular problem of clinical concept recognition when dealing with Portuguese natural language text. After consulting with several MD (Medical Doctors) whose activity is the main subject of representing knowledge that way, we found that we can take into our advantage the fact that each one usually develops his/her own way of writing down their daily chores. What we have developed is a way of maintaining an acquired controlled vocabulary. Assuming that this is a task involving NLP (Natural Language Processing) specifically for Portuguese we tried to develop some techniques that can be easily applicable to different languages with the same set of constraints presented ahead.

2 Experiences in the field

2.1 The ULSNA experiment

The ULSNA, E.P.E. (http://www.ulsna.min-saude.pt/) has as its principal object the provision of primary and secondary health care, rehabilitation, palliative and integrated continued care to the population and the means necessary to exercise the powers of the health authority in the geographic area affected by it. ULSNA is a healthcare providing regional system that includes 2 hospitals (José Maria Grande in Portalegre and Santa Luzia in Elvas) and the primary care centers in all the district counties. Universidade de Évora signed an agreement with ULSNA that enabled the usage of de-identified (according to safe-harbor principles as reviewed in [3]) clinical data from the SAM system in use both in the Primary Healthcare units and in the Hospitals. Using the clinical data that is available for us we intend to take advantage of the tooling available to reach the objectives mentioned in section 1.

2.2 Inevitability of English usage in Bioinformatics

In the Biomedical informatics domain the knowledge representation of choice has been evolving since the initial developments of the Gene Ontology project (http://www.geneontology.org/) in the beginning of the century into ontology based representations. With the advent of research about reasoning capabilities and its applicability in the Semantic Web a lot of services that expose that capabilities have appeared but all of them, that matter, exploring only English based terminology. Recently we have been testifying the appearance of a strong set of Web Services in the area of semantic annotation of texts in the Biomedical domain. These specialized, domain oriented services, are made available mainly by the U.S. NCBO (National Center for Biomedical Ontologies) through BioPortal (http://bioportal.bioontology.org/) or by the NLM (National Library of Medicine) itself through the UTS (Unified Terminology Services) interface. We can find among these, services that do all kind of text processing for an information extraction pipeline like POS (Part of Speech) tagging, NER (Named Entity Recognition), and for clinical concepts both CSD (Concept Sense Disambiguation) and SSE (Semantic Similarity Estimation) for instance. To be able to fully
exploit these in our work we definitely need to translate very precisely from the
Portuguese personal medical jargon to an English completely understandable by
the annotating services.

3 How to represent an acquired controlled vocabulary

3.1 Translation Memory as a jargon repository

In the world of professional CAT, Computer Aided Translation, the use of per-
sonal terminologies has been subject to standardization using the TM [9], trans-
lation memory, model. A TM, is a database that stores so-called "segments",
which can be acronyms, nicknames, words, phrases and paragraphs that have
already been translated, in order to aid human translators through the CAT
software. The source text and its corresponding translation in language pairs
called “translation units” are stored in the TM in a standardized exchangeable
form.

3.2 TMM1 tooling

From the huge listing of resources available currently in the internet we checked
carefully those that can handle TM2 based in the TMX Standard. TMX was
originally developed and maintained by OSCAR (Open Standards for Conta-
iner/Content Allowing Re-use) a special interest group of LISA (Localization
Industry Standards Association). The format allows easier exchange of transla-
tion memory between tools and/or translators with little or no loss of critical
data. TMM, Translation Memory Managers, that handle creation and main te-
nance of the TMX files can be desktop centric or centralized TM systems that
store on a central server. They work together with desktop TMM and can in-
crease TM match rates by 30-60% more than the TM leverage attained by desk-
Top TMM alone. They can very fruitfully develop Machine Translation based TM
that can then be exported and further refined by the desktop systems. We ex-
lored 2 systems that proved to be interoperable given the standardized formats
that are handled: Google translator toolkit (https://translate.google.com/toolkit/) and
mymemory translation services (http://mymemory.translated.net/) that
will be detailed further with an appropriate example.

3.3 Structure and Extensibility

The first step is the creation of a seminal TM for some sample documents
by using MT3. Clearly all the acronyms and personal defined nicks were not
matched but we can then download a base TM to be further refined. TM
files have an incremental XML structure that can be very easily manipulated

---

1 Translation Memory Manager
2 Translation Memories
3 Machine Translation
trough adequate XSLT transforms, that can be manually applied or enqueued in an automated workflow as we show in section 5. However, some tools nurtured our possibilities and rendered our work even simpler. Those are the open source OmegaT (http://www.omegat.org/en/resources.html/) and its companion the Okapi framework (http://okapi.opentag.com/). For every translation job, OmegaT creates a set of project folders which contain or can contain specific files. With appropriate file positioning, as can be learned from the OmegaT documentation, one can manipulate the sources and translations. We use the file positioning capabilities of our machine introduced in section 5 to automate all TM management. The knowledge accumulation for our specific translation tasks are just the positioning of the subsequent developed TM in the specified folder. We have then a TM enrichment workflow that can incrementally match more and more concepts. TMs can even be orchestrated by service or specialty for example rendering a fewer amount of work left to be done individually.

4 Automated acquisition from Clinical Episodes Text

As reviewed by the authors in [2] the state-of-the-art for acquisition from Clinical Text has enjoyed strong developments in recent years. In the mentioned paper we present a proposal for automated acquisition from HL7 messaging but, here we are delving into the more generic possibility of extracting from free text present in most interfaces used by clinicians. Going from clinical episodes free text that is usually presented in a human friendly format to one adequate for computer processing involves a fair amount of text processing to handle different situations because: (1) Reports aggregate information from different clinical episodes that are not uniquely identified or not even individually dated, (2) The clinician is only identified by his/her name if any identification is made at all, (3) The information conveyed in free text is intended only to be understandable by fellow practitioners or even by the clinician him/herself making use of pragmatic jargon normally plagued with acronyms and nicknames abundant in their specific community, (4) Text is profoundly intermixed with decorative elements for better legibility, normally in PDF or HTML files, (5) The clinicians natural language is other than English without concepts defined in a foundational thesaurus like SNOMED CT or FMA (Foundational Model of Anatomy) for instance that don’t exist in that particular language, (6) The time spanning of the processes depicted in natural language is very difficult to represent formally.

4.1 The adequate annotation workflow

A set of sequential steps must be used to go from the pure text to the extracted concept instance. Those steps workflow can be configured declaratively using the software architecture shown in section 5. The translation steps involved are: (1) Manual translation (that is indispensable for the translator tutoring) with the precise clinicians validation of their jargon adequately translated into English, (2) PDF to raw text, or to structured (XML), converting for adequate documents
cleansing. The remote annotation steps are: (3) NER (Named Entity Recognition) of all the patient names, clinician names, addresses, symptoms, signs and prescriptions with all the acronyms, units, time-spanning and time-snapping involved with the usual short forms that a particular doctor usually uses. In our particular situation we maintain a local cache of the previously identified vocabulary to check exact concept matching, (4) WSD (Word Sense Disambiguation) where terms can be disambiguated without technical clinical expertise. Most of them however have to be disambiguated according to the previously acquired concepts in our controlled vocabulary, (5) EAV (Extraction of Attributes and Values) is the final pure, single language, task that has to be performed and in which we need the tooling that this paper refers to filter the concepts from the annotated text to extract concept instances, (6) Semantic annotation using the Web interface of any of the services that we introduce in section 4.2, either manually using the interactive interfaces or automatically with the Web Services available.

Given the array of Web Services that can semantically annotate bio-medical concepts in English in section 4.2, we chose to use an evolutionary approach for use of the BioPortal annotator. We first use the annotator after manual preprocessing for the TM tutoring and later a fully automatic workflow based in Web Services orchestration.

4.2 SAM Corpus

The number of registered clinical episodes from which we can extract our information induces any ontology or knowledge engineer in a very hard problem to solve just trying to figure out the granularity of the ontology. When trying to apply the principles of well defined formal ontologies depicted in [8] and trying to avoid the errors mentioned in [1] we decided to get our hands wet with a simple approach to the representation of disease and diagnostic illustrated in [7]. In our particular case we face a shortage of structure of the reports in SAM in order to extract the intended information in a systematic way. We illustrate below that no complete demographic information is available in any of the used reports or that the problems enumerated in the Problem List don’t have any kind of severity or progress information associated with them.

The SOAP report is the most informative of all available as it depicts a clinical encounter in a semi-structured way in a datable point of time. The Subjective, Objective, Assessment, Plan (SOAP) framework, used to structure progress notes to facilitate problem specific, clinical decision making by physicians, is a well-known, canonical structure in the medical domain. The underlying structure of the SOAP report induces some very important assumptions. We find sections that can be associated with Subjective, the symptoms section S where we extract directly into a cpr:symptom instance, medications found here are only those administered during the patient visit. Objective, the O section where are sign records that we take as generator for cpr:clinical-findings. Assessment, the analysis section A which are the clinical investigation acts whose outputs can be clinical artifacts to investigate things that can be consequence of any
of physiological or pathological processes. Finally Plan, the P section where the therapeutic acts can be extracted with all the timing, posology and prescriptions registered in that particular clinical encounter, medications here are prescribed for discharge [6]. We now take advantage of the fact that we have to translate from jargon to English to customize our centralized TMM like the Google translator toolkit or mymemory translation services enhanced with our own Translation Memories and Glossaries. Let us introduce some demonstrative examples taken from a sample document gently provided by Dr. Carlos Baeta and properly de-identified:

We will, in the process of using one of the services, create Translation Memories with the identified personal acronyms like:

AP (Antecedentes Pessoais) into Personal History, HTA (Hiper Tensão Arterial) into High Blood Pressure, FA (Fibrilhação Auricular) into Atrial Fibrillation and V. Mitral (Válvula Mitral) into Mitral Valve. Some of which are acronyms that can be given the suitable translated concept like: ECG (Electro Cardio Grama) into Electro Cardio Gram or those that are even English acronyms them-
selves like INR (International Normalized Ratio) into International Normalized Ratio and some which are not really needed because the conveyed information is irrespective of what language is in, like: CENTRO DE SAÚDE into HEALTH CENTER or SEDE into MAIN OFFICE.

Included in this sample are notorious some more complex problems that are not related to the translation itself but with some other problems like the time spanning of concepts like “1 comp/dia” which is adequately translated to “1 tablet per day” using the defined Translation Memory but has to be posteriorly well defined as a time delimited occurring process in the process of CPR ontology enrichment.

Services to annotate clinical concepts in free text - Apart from being able to provide “our own” Web Services for various tasks given the availability of downloading several types of terminologies like MeSH or SNOMED CT CORE and generally the UMLS Metathesaurus, currently there are a myriad of WS at reach that can be configured to be connected to our CP-ESB presented in section 5 as service providers. Among those we think that are worth mentioned here the BIOPortal, OntoCAT (http://www.ontocat.org/) and UTS.

All these provide Web Services that offer specific tasks for Biomedicine terminology. Carefully chosen endpoints provide features that range from simple term lookups to complete semantic concept acquisition. With the concepts identified in our, purpose directed, controlled vocabulary its only a matter of direct translation with the aid of the mentioned TM to be able to semantically annotate.

Complex Clinical Concepts - The formal way that CPR was precisely defined and is expected to be populated urges us to handle some more complex problems like time-spanning, time-snapping of processes or the anatomical location of the collected phenomena using the suggested Ontological Realism. Our approach is to just extract the instances and populate according to the suggestions in [7] and leave all the inferring possibilities to the OWL DL reasoners and interested clinicians. Although more comprehensively explored by us in [3] we are collecting essentially cpr:medical-problems and to render evident the complexity of the population task and the CPR ontology as an illustrative example we can see how a therapeutical prescription instances a span:ProcessualEntity that cpr:occursDuring a given dateTime and we have to cleverly delimit it’s beginning as cpr:startsNoEarlierThen and cpr:startsNoLaterThen and its conclusion between cpr:stopsNoEarlierThen and cpr:stopsNoLaterThen. A significant number of this kind of "smart thinking" commitments has to be done to embrace notions of snapping (like medication administration obviously) or anatomical placement where the FMA is used as a feeder/validator ontology. We use all the alignment of ontologies used to form the CPR as both feeding as well as validating models in the process of instance extraction from the texts.

These are: FMA for anatomy, OWL Time for temporal relations, BioTop for physics and structural/systems biology, BFO for the ontology philosophy, SNOMED CT for the clinical documentation and reporting and finally the Relations Ontology that formalizes all the relations. Very important is the
use of coding of all of the occurring episodes according to the SNOMED CT standard that will allow our representations to be very widely interoperable.

5 Software Architecture

We have presented the full conception of an architecture in [3]. The CP-ESP (Clinical Practice - Enterprise Service Bus) is a common rail where messaging can flow using a subscription model that enables the communication to be detached from any two particular services but instead be available on-request by one and served by another in a loosely coupled way. The ESB can then intervene in message exchange and overwrite standard rules for service execution. The case of an intervention here is the ability to filter and redirect invocations to the appropriate NLP task processors depending on the source being labeled with the status of the load it carries. The REST philosophy is suggested in our proposal as the best way of implementing a Service Oriented Architecture that serves as the communication underlying structure of our system. REST endpoints are available for the generality of our needs. The available Web Services can render responses in highly-structured forms like JSON or in any of the standardized mime types that can be handled by the filtering and queuing capabilities of any configurable available ESB like those based in Apache ServiceMix (http://servicemix.apache.org/home.html) or Mule (http://www.mulesoft.org/) for instance. They can be configured to compose a complete pipeline very easily:

![TM Workflow with CP-ESB](image)

**Fig. 2. TM Creation Workflow**

In the picture is shown the possibility of defining services (daemons) that monitor the presence of new reports and appropriately apply the needed transformations according to their status and content until a translated document is
delivered to the adequate end-point for annotation. All of this can even be done with due care about scalability, availability and all problems associated with a modern state-of-the-art software architecture as presented in [3].

In the picture is shown the possibility of defining services (daemons) that monitor the presence of new reports and appropriately apply the needed transformations according to their status and content until a translated document is delivered to the adequate end-point for annotation. All of this can even be done with due care about scalability, availability and all problems associated with a modern state-of-the-art software architecture as presented in [3]. Building over the suggested infrastructure the systems are rather composed as opposed to monolithically built and so manifest high capabilities of plug-and-play configuration allowing for interchangeable providers (as Web Services), Reference Ontologies (Feeders), and target ontologies. Having the foundations available with the right weapons provided one has to take a practical approach to the development of a target system using, in our case, the Java best-practices for pragmatic development that include a number of Patterns as in JEE (Java Enterprise Edition) or the pragmatic approaches developed in such successful projects as Spring (http://www.springsource.com/).

The flowchart that depicts graphically the acquisition from the source texts in Portuguese to the creation of the appropriate CPR instance is:

![Acquisition Flowchart](http://example.com/acquisition_flowchart.png)
6 Future Work

In 2007 the work done under the W3C - Semantic Web Healthcare and Life Sciences Interest Group auspices rendered amazing contributions done with the most interesting technological landscape available at the time. Some realities that have since evolved and we could use for the present work are:

- The POMR Ontology that was started to be developed in 2006 was the target of the fruitful GRDDL efforts that appeared during 2007 in which base we are now working on. This efforts however were rdfs based for the OWL2 developments were still incipient. Only in November 2009 W3C standardized a OWL based ontology called the CPR that we use in our work.
- For a given HL7 CDA document the GRDDL gleaning presented in 2007 with it associate transform creates one cpr v0.5 instance with one cpr:patient-record with as many cpr:screening-acts and associated cpr:clinical-descriptions as episodes are referred in the CDA source. This approach, though impressive technology, lacks the granularity that can be rendered now with further developments that CPR has gained in its latest versions.

Taking advantage of the developments done so far we intend to build upon what was then achieved and propose work to be developed like:

6.1 Extend our proposal to use GRDDL for gleaning HL7 documents into CPR

The same architecture that we are proposing for gleaning from BioPortal’s OBA output may be, as well, used to extract resources from CDA documents compliant with RIM V3. However it seems obvious that the complexity that the specification conforms could easily “hinder the trees and not show the forest”. Many recent proposals are being developed to apply restricted templates to make RIM tractable. The most notorious, that we intend to use here, is the greenCDA project that already has delivered results. In fact, the greenCDA specification has been officially balloted via HL7 and the version 1.0 is already available for download.

6.2 Restrict CPR to be OWL DL well formed or even further

Unfortunately, CPR is not currently OWL DL syntax conformant. For the current reasoners to work effectively with OWL based ontologies they have to be OWL DL to be validatable in linear time. Although our clinical practice ontologies are restricted in their size most of the times, the reasoners check the validity a-priori and tend to refuse to work when some non-conformities are found. One interesting line of research may be the adequate restriction of CPR to be OWL DL conformant without lost of the needed expressiveness for clinical practice representation. We found even arguable that with adequate restrictions, like elimination of transitivity [3] a Horn SHIQ ontology can be developed from
CPR and then we reach the possibility of applying Consequence-Driven Reasoners that have been shown recently to be very attractive for dealing with huge data-sets.

6.3 Pre curating of the text terminology for clinical concept orientation

One of the main issues that were observable during the trial tests was that it is crucial to orientate the texts written in the SOAP reports for achieving a more direct clinical concept acquisition from SNOMED CT or CORE. When using the NCBO annotator even with only one level of expansion all the duplicate, obsolete and under-adequate terms were found and this rendered the need for a post annotation manual refinement task. This is incompatible with the full automation proposal that we intend to provide here. The solution then is to have local SNOMED CT validation based in AJAX completion to provide a-priory correct terminology for post annotation term adequacy.

7 Results and Conclusion

While the process of generating personal TMs has been developing simultaneously with other CPR population tasks the initial impressions look promising. The adequacy of the method and tools proposed here, however, are to be evaluated by the clinicians themselves. We are currently in the process of assessing the manual processes with a selected group that will after developing their own TMs, at their own pace, integrate their TMs in the complete ontology acquisition workflow to be able to assess the reasoning capabilities that can be made available. We are developing a process of quality assessment that will involve questionnaires that will be presented as soon as available and processed, expected during 2012. After this controlled group we expect that the generated TMs can be included in a regional corpus that will ease extremely the adherence of the rest of the community and try to expand to regional level which shall then be a straightforward step to achieve.

We demonstrate the articulation needed of different software tools, as well as individual medical knowledge, to extract the individual clinical concepts to develop personal TMs for automatic alignment of Portuguese texts to English ontologies, thesaurus and controlled vocabularies. The texts used are particular MD practice reports with instances collected automatically from their usage of the specific local SAM system. This proposed methodology and architecture is applicable to whatever EHR that produces clinical reports in whatever language in the world so it should be easily applicable globally.
Bibliography


