Basics of a Drug Ontology for Annotations of Clinical Narratives

Zdenko KASÁČ[a,1], Catalina MARTÍNEZ-COSTA[a], Markus KREUZTHALER[a,b] and Stefan SCHULZ[a]

[a] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*
[b] *CBmed GmbH - Center for Biomarker Research in Medicine, Graz, Austria*

Abstract. Although the medication given to a patient during a hospital stay is fundamental for patient safety, quality assessment and economic aspects of health care services, in-patient drug prescriptions are still commonly done on paper in many institutions. In Europe and North America most of the pharmacies are computerized. This does not guarantee, however, that drug prescription information is always available in hospital information systems in a structured way. As long as the only computer-readable sources of medication information are clinical narratives such as dictated or typed discharge or outpatient letters, analysis of drug information depends on natural language processing (NLP). The performance of NLP critically depends on annotated clinical corpora. We are currently developing an annotation schema for mentions of drugs in discharge letters, of which ontological foundations are presented.

Keywords. Annotations, clinical narratives, drugs and substances, formal ontology

1. Introduction

Manually annotated texts are key resources for information extraction systems. However, if the underlying annotation schema does not offer clear-cut criteria, a high agreement between human annotators is hard to achieve. The avoidance of ambiguity is therefore a prerequisite for precise human and machine annotations. We believe that definitions based on formal ontologies are helpful for the provision of precise annotation criteria.

The documentation of drugs given to a patient during a hospital stay is fundamental for patient safety, quality assessment and economic aspects of health care services. Nevertheless drug prescriptions are still commonly done on paper in many institutions. In Europe and North America most of the pharmacies are computerized, but this does not guarantee that prescription information is always available in hospital information systems in a structured format.

The primary purpose of clinical narratives in electronic health records is communication between clinicians. These texts are written to be naturally understandable, ideally across specialties. The mention of drugs in a clinical document is, however, not limited to the prescription section. Drug names can also be found as

[1] Corresponding author: Zdenko Kasáč, Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria; E-mail: zdenko.kasac@medunigraz.at

mentions of drug allergies, suspected triggers of side effects, drugs discussed but not prescribed, drugs suspended etc.

The mention of drugs names alone is not yet sufficient. It is often be supplemented by strength, route, dose form and regimen, and it is embedded in a context, such as affirmation, negation, or intentionality. In this paper, we focus only on text passages that refer to drugs in a strict sense, not on the text that provides additional context, although our future goal is to provide a full picture of drug mentions in clinical documents. A fully-fledged ontology should also formalise drug and product families, categorise names and their parts, introduce the notions of branded and packaged drugs and, finally, provide formalisations of signature information, including drug strengths, dose forms and dosing regimens, routes and temporal information. Finally such an ontology should be compliant with existing and emerging standards, such as ISO IDMP[1] and Standard Terms[1].

2. Methods

First analyses of clinical summary samples were done to shape the ontological definitions, in parallel with the bottom-up formulation of a drug annotation guideline. In our text samples, most textual references to drugs were found within the prescription section of clinical summaries. It is expected that the patient's GPs base their medication decisions on these recommendations, and the assumption is reasonable that the drugs mentioned in the prescription section were also administered during the hospital stay and, possibly, even before hospitalisation. It is a known problem in our setting (an Austrian university hospital) that drug therapies initiated and finished during an inpatient stay are often not represented in computer-readable EHR documents.

We analyse drug mentions in clinical documents from both an ontology and language engineering point of view. Formal ontologies state what is universally true in a domain, using logical axioms. Upper-level ontologies guide the ontology engineering process by supporting ontology engineers with a consistent framework, which can foster semantic interoperability if used by different domain ontologies. Our drug model is aligned with two existing ontologies, viz. the domain upper-level ontology BTL2 [1] and the clinical ontology SNOMED CT [2]. Our current work is related with earlier work inspired by the FHIR medication resources and their ontological interpretation [3], created in a bottom-up way from clinical summaries.

The basis of our analysis is a corpus of pseudonymised discharge summaries filtered by the diagnosis Malignant Melanoma (ICD-10:C43), counting 400 documents. The annotation workflow was defined as follows: mentions of drugs, signatures, etc., were read and annotated by a medical doctor. Subsequently, first 200 of these were re-read and checked for annotation errors by the same annotator. Afterwards, the contained annotations were compared to machine-generated annotations that were created using a drug dictionary. All conflicts of annotations were checked by the human annotator and mistakes detected during the crosscheck were manually corrected by the same person.

The aim of the work was to identify semantic categories (classes), which upon annotation can be related to text tokens in the clinical narrative. The purpose is to fill a structured dataset in a way that all the relevant parameters of a prescription are

represented. The following priorities, or guidelines, steered the bottom-up evolution of the annotation schema:

- If a problem in the schema exists and requires a change of the model, the change that results in the lowest complexity is preferred.
- Any change, especially if it increases complexity, must be assessed by its practical usefulness.
- The goal is to provide a semantic structure that is simple, yet precise and potentially applicable of narratives from other sources.

Based on the annotation experience, we created a model of semantic annotation. It is centred on Drug annotation, composed of three semantic categories: *DrugProduct*, *DrugSubstance* and *DrugFamily*. Drug is defined as a "chemical substance used in the treatment, cure, prevention, or diagnosis of disease or used to otherwise enhance physical or mental well-being"[4]. As the term "drug" is often also used for packaged and branded preparations, we use it only in the following, refined way:

- *DrugSubstance* classifies amounts of matter considered homogeneous, ideally constituted by a single, ideally chemically defined active ingredients.
- *DrugProduct* identifies a specific type of preparation, typically denoted by a brand name. Alternatively it can also be an individually compounded medication.
- *DrugFamily* is a meta-attribute that the annotated token(s) denote a superclass of drug substances.

In the following OWL DL [5] expressions we refer to SNOMED CT by the prefix "sct" and to BioTopLite2 (BTL2) by the prefix "btl2". We use OWL Manchester syntax. OWL classes are represented in Italics and OWL relations (object properties) in bold face. We first analyse the mention of the drug itself, for which we distinguish between drug substances, drug families and drug names.

Specific drug substances can be hierarchically ordered. For instance, *Penicillin G* is a *Penicillin*, which is a *BetaLactam Antibiotic*, which is an *Antibiotic*, which is a *DrugSubstance*:

PenicillinG subclassOf *Penicillin*

Penicillin subclassOf *BetaLactamAntibiotic*

BetaLactamAntibiotic subclassOf *Antibiotic*

Antibiotic subclassOf *sct:Substance*

In our ontology approach, which does not allow for meta-categories, *DrugFamily* does not appear as a separate category, but just as a way to flag those nodes in the drug substance hierarchy that are not terminal. One way to conceive drug families is to express them as information object classes, with substance names as instances. These substance names then extend to ontology classes in the substance hierarchy. The mention of just drug categories would be incomplete in the prescription section, but it is rather common in other parts, e.g. in passages like "treated with low dose non-selective NSAIDs under PPI protection", or "we kindly ask a specialist to optimize the cardiac medication".

Ontologically, *DrugSubstance* and *DrugProduct* are strictly different. In SNOMED CT, substance concepts extend to amounts of pure or mixed substances, whereas drug product concepts extend to countable entities like pills, which include active and inactive ingredients with a defined shape, strength and often even manufacturer. In our formalisation, following SNOMED CT, there is a logical implication between a drug product and one or more substances. If this information is complete, then a drug product always allows to infer its active ingredients but not the other way around:

sct:Diclofenac(product) subclassOf
sct:hasActiveIngredient sct:Diclofenac(substance)

3. Preliminary observations and conclusions

In the following, the term "drug" is used for branded products, active ingredients and drug families. The analysis showed where these are mentioned within the narratives.

- *Diagnosis section*. High impact drugs are mentioned together with the diagnoses. Typically, past and present cycles of chemotherapy are mentioned here, as well as the status of oral anticoagulation, etc.
- *Radiology section*. Injected contrast agents are usually mentioned.
- *Lab result section*. Occasionally, substances are mentioned in the lab results as entry points for therapeutic drug monitoring (TDM).
- *Prescription section*. Commonly, most medications in a clinical summary are mentioned here in a semi-structured form, mostly as branded drugs with signature, occasionally also as recipes for drug compounding.
- *Across the document*. Elaborations regarding medications to be changed, indications thereof, as well as medications like local anaesthetics that were directly applied are mentioned in several parts, mainly the last section of the document.

From an ontology point of view we now scrutinize the textual source, its denoting entities and especially the way of referencing to reality. Medical texts, as well as parts thereof are information objects in the sense of BTL2 (a subclass of 'generically dependent continuant' in BFO).

Text subclassOf *btl2:InformationObject*

DenotingEntity subclassOf *btl2:InformationObject* and **btl2:isPartOf** some *Text*

We introduce the classes *MedicationAdministration* and *MedicationPrescription* as subclasses of *btl2:Action*:

MedicationAdministration subClassOf *Action* and
btl2:hasAgent some *Human* and **btl2:hasPatient** some *DrugProduct*

MedicationPrescription equivalentTo *Plan* and
btl2:hasRealization only *MedicationAdministration*

As mentioned, the fact that there is a drug prescription issued at the end of a hospital stay implies that the drug was most probably administered during the stay. A drug prescription could therefore also be read as a (highly probable) drug administration.

Let us assume **e** is a denoting entity within a text **t**, regarding prescription of drug D. The double meaning can then be expressed as follows:

e Type (*DenotingEntity* and **btl2:hasPart** some *CertainInformation* and **btl2:represents** only *MedicationPrescription* and **btl2:hasRealization** only (*MedicationAdministration* and **btl2:hasPatient** some *D*)

OR

e Type (*DenotingEntity* and **btl2:hasPart** some *ProbableInformation* and **btl2:represents** only (*MedicationAdministration* and **btl2:hasPatient** some *D*))

We now turn briefly to the annotation process and the elaboration of the annotation schema. The annotation has to anticipate that the denoting entity does not always refer to a *DrugProduct* class. In this case the annotation is done with a subclass of *DrugSubstance* at the leaf node level as well as at a more generic (*DrugFamily*) level. The fact that drug mentions in clinical summaries are result from human writers or dictated and subsequently voice-recognized or human-typed, several challenges arise:

- Misspellings or typing errors
- Nominal anaphora: a drug product is mentioned in some part of text; the same referent is mentioned by a more general expression in the following text
- Ellipsis: omission of information, such as strength, where it is already known, obvious or irrelevant
- Convenience mixtures of names of branded drugs and ingredients
- Abbreviations, e.g. "LA" for "local anaesthesia"
- Detailed prescriptions for individual drug compounding
- Detailed, unstructured, free-text instructions

Table 1. Drug related annotation labels with their typical occurrences in the corpus of German clinical narratives under investigation.

Annotation labels	Examples	Definition and annotation instructions
DrugProduct (subclass of SNOMED CT "Pharmaceutical / biologic product (product)") Primarily industrially manufactured and registered drugs, sometimes traditional individual compounding.	*"Concor COR"* *"Trombo Ass"* *"100ml NaCl"* ("100ml" is rather exceptionally part of the token that belongs to this category because it specifies the product used, not merely an amount applied) *"Roferon", "Selocen plus retard"*	Defines preparation name and/or brand and active substance of the drug. A textual expression annotated as DrugProduct stands for a particular brand name or name of a compounded drug.
DrugSubstance (subclass of SNOMED CT "Substance (substance)") States the name of the active ingredient.	*"Bisoprolol"* *"C_2H_5OH"* *"DTIC"* *"Carboplatin"*	A textual expression annotated as DrugSubstance represents a substance; this substance may occur in many brand names.
DrugFamily (non-terminal subclass of SNOMED CT "Substance (substance)" or "Pharmaceutical / biologic product (product)") Denotes pharmacological class or a group of medications delimited by other means.	"*LA*" (local anaesthesia) "*Herzmedikation*" (Cardiac medications) "Antibiose" (antibiotics) "*Dauerinfektionsprophylaxe*" (prophylactic antibiotic therapy) "*Schmerztherapie*" (pain therapy)	DrugFamily aggregates agents (by indication such as antibiotics, antidepressants, by mechanism such as β -blocker, ACEi) or by other criteria.

The core annotation schema provided in Table 1 poses several challenges for the completion of the drug prescription ontology. In the present abstract, we intentionally omit other parts of the schema that deliver specific temporal, logical and signature information to a given drug mention within a narrative. Finally, the representation of conditions that act as decision points must extend the scope to observables and clinical findings. The necessary formalizations would probably go beyond what can be represented by OWL-DL description logics. The same might be true with the interpretation of drug mentions in the context of therapeutic drug monitoring, which has been, so far, excluded from our annotation schema.

Boundary issues may arise like the therapy with submolecular particles (e.g. gamma knife, radiotherapy, and phototherapy). Its characterisation as drug therapy might be contentious. Similar considerations apply to functional foods and special diets as a category between drugs and foods.

References

[1] Schulz, S., Boeker, M., & Martinez-Costa, C. (2017). The BioTop Family of Upper Level Ontological Resources for Biomedicine. Studies in health technology and informatics, 235, 441.

[2] SNOMED International. http://www.snomed.org/snomed-ct

[3] Martinez-Costa, C., & Schulz, S. (2017). HL7 FHIR: Ontological Reinterpretation of Medication Resources. Studies in health technology and informatics, 235, 451.

[4] Dictionary.com Unabridged. Based on the Random House Dictionary, © Random House, Inc. 2017. http://www.dictionary.com/browse/drug

[5] Baader, F. et al., *The Description Logic Handbook,* Cambridge University Press, Cambridge, 2007 OWL 2 Manchester syntax https://www.w3.org/TR/owl2-manchester-syntax/

[6] IDMP Standards - Identification of Medicinal Products. (n.d.). Retrieved November 10, 2017, from https://www.idmp1.com/

[7] Standard Terms. (n.d.). Retrieved November 10, 2017, from https://standardterms.edqm.eu/