Requirements for natural language understanding in electronic health record systems based on the paradigm of referent tracking

(Extended abstract)

Werner Ceusters (1,4), Barry Smith (2,3,4)

(1) European Centre for Ontological Research, Saarbrücken, Germany

(2) Institute for Formal Ontology and Medical Information Science, Saarbrücken, Germany

(3) Department of Philosophy, University at Buffalo, NY, USA

(4) National Center for Biomedical Ontology, University at Buffalo, NY, USA

1 Introduction

Electronic health records (EHRs) consist primarily of descriptions of a patient's medical condition, the treatments administered, and the outcomes obtained. These descriptions are about concrete entities in reality: for example about the particular pain that the particular patient John experienced in his chest on this specific day; or about the particular pacemaker – with its specific serial number assigned to it by its manufacturer – that was implanted in John during the particular surgical procedure that started at a precise moment in time on a certain day.

The descriptions contained in current EHRs contain very few explicit references to such entities. This lack of explicit reference is usually a minor problem for human interpreters, but it makes an accurate understanding of EHR data nearly impossible for machines. This is because reference resolution in running text (still the most common format for descriptions in EHRs) is one of the hardest problems in natural language understanding [1]. But even those EHR systems which incorporate data in more structured formats, for example by resorting to controlled vocabularies, terminologies or ontologies, are in no better shape in this respect. This is because the terms or codes contained in the latter are used simply as an alternative to what would otherwise have been registered by means of general terms in natural language. By picking a code from such a system and then registering that code in an EHR, one refers generically to *some* instance of the class represented by the code. It is still left at best only partially, and indirectly, specified which particular instance is intended in concrete reality.

We therefore argue that, where coding systems and terminologies provide rich vocabularies to describe the entities that exist in reality using general terms, there is a need also for an associated mechanism for expressing *what those descriptions are about*, i.e. the particular entities in reality to which they refer. Such a mechanism is indispensable at the interface

where coding systems meet the clinical record if we are to gain maximal advantage from coding efforts and from so-called formal representations and descriptions in EHR systems [2].

2 Referent tracking

Drawing upon our experience in EHR research and standardisation [3, 4, 5], and also from our philosophical research on universals and particulars [6], we introduced referent tracking as a paradigm under which it will become possible to refer explicitly to all of the concrete individual entities relevant to the accurate description of each patient's condition, therapies, and outcomes through the assignment of unique identifiers [7, 8]. Such an identifier is called a IUI (pronounced to rhyme with 'CUI' as used in the UMLS [9]), for 'Instance Unique Identifier'. This means that, not only does each patient receive a IUI, but so also does the particular fracture he is suffering from, the particular bone that is fractured, and even, if the clinician finds this important, the particular pain the patient is experiencing in a certain time period, or the particular document in which the pain is first recorded.

As such, referent tracking goes much further than current practices, under which entities are uniquely identified only when they belong to a restricted range of entity-types, including human beings (the patient himself, the physicians involved), buildings (the hospital in which the patient is treated), certain instruments and devices, and so forth. Moreover, where the majority of entities uniquely identified under current schemes are outside the patient (physicians, instruments, wards, X-ray images), referent tracking extends the facility of unique identification also to the patient's body parts, the specific diseases he has suffered from, the symptoms he has exhibited, and so forth. It goes beyond established approaches also in the degree to which it takes seriously the notion of 'uniqueness'. For where, in many EHR systems, patients and physicians are uniquely identified only relative to some local context (for example of the hospital in which a given EHR-system is used), referent tracking aims for *global* uniqueness.

Note that IUIs refer to the real entities themselves out there in reality, and not to data about these entities. IUIs are the means whereby those constellations of particular entities (tokens, instances) in reality that are relevant to clinical care can be represented in an EHR in the same direct way in which the corresponding classes (types, universals) are already represented by means of clinical coding systems.

Thus IUIs are also *not the entities themselves*. This might seem obvious, but use-mention confusions ('Swimming is healthy and is a concept included in our terminology') – in which an entity in reality and its representation are confounded together – are abundantly present in the literature on knowledge representation in general and on concept-based terminology systems in particular [10].

3 Statements in referent tracking systems

In [8] we explored ways in which the referent tracking paradigm can be implemented in the healthcare environment. Our hypothesis is that, once the right infrastructure is in place, the burden on clinicians and nurses (or on whomever is assigned the task of registering patient data) will be not significantly greater than under existing strategies for data entry – but that the benefits, in terms of semantic interoperability of computer systems and also in terms of patient management, cost containment, epidemiology and disease control, as well as for the advance of science in the domain of biomedicine, can be enormous.

The purpose of a referent tracking system (RTS) is, as its name suggests, to keep track of *referents*. Referents are entities that exist in reality, i.e. in the spatiotemporal world that surrounds us. Most referents are *particulars*, examples being a copy of the manual in which this paper is published, and its authors. Other referents are *universals*, examples being *journal*, *manual*, *person*, H_2O. and so forth.

An RTS will primarily contain information about particulars. The users who enter this information will be required to employ IUIs in order to assure explicit reference to the particulars about which they are providing information. Thus the information that is currently captured in the EHR by means of sentences such as: "this patient has a left elbow fracture", would in the future be conveyed by means of descriptions such as "#IUI-5089 is located in #IUI-7120", together with associated information to the effect that "IUI-7120" refers to the patient under scrutiny, and "IUI-5089" to a particular fracture in patient #IUI-7120 (and not to some similar left elbow fracture from which he suffered earlier). The RTS must correspondingly contain information relating particulars to universals, such as "IUI-5089 **instance_of** fracture" (where 'fracture' might be replaced by a unique identifier pointing to the representation of the universal *fracture* in an ontology) [6].

4 Natural language processing for referent tracking

Of course, EHR systems that endorse the referent tracking paradigm should have mechanisms to capture such information in an easy and intuitive way, including mechanisms to translate generic statements into the intended concrete form, a form which may itself be operative primarily behind the scenes, so that the IUIs themselves remain invisible to the human user. One could indeed imagine that natural language processing (NLP) software will one day be in a position to replace in a reliable fashion the generic terms in a sentence ('John's mother', 'John's pacemaker') with corresponding IUIs for the particulars thereby denoted, with manual support in flagged problematic cases. This corresponds on the level of particulars to what users already expect from EHR systems on the level of universals in supporting entry of codes or terms from coding systems.

The requirements for such natural language analysers are thus:

- to distinguish in clinical narrative the words and phrases that refer to either particulars or universals or (as will commonly be the case) both;
- to identify what specific particulars (among those already described in the RTS) are referred to by the terms and phrases that denote particulars, and what universals (in terms of the foundational ontology and associated domain ontologies linked to the RTS) are referred to by the terms and phrases that denote universals;
- to identify those particulars and universals not thus far recognised within the RTS or the associated ontologies respectively, and to incorporate them appropriately into the system;
- to create adequate linguistic ontologies and link these to foundational ontologies in a way that will enable the machine to infer from the semantic (as understood in linguistics) relationships between the words and phrases in a source text what the ontological relationships are that obtain between the particulars and universals referred to;
- to represent the results of this analysis in such a way that semi-automated population of the RTS is achievable.

References

[1] Popescu-Belis A and Lalanne D. Reference resolution over a restricted domain: References to documents. ACL 2004 Workshop on Reference Resolution and its Applications, Barcelona, Spain, July 2004;:71-78.

[2] Smith B, Ceusters W. An Ontology-Based Methodology for the Migration of Biomedical Terminologies to Electronic Health Records. AMIA 2005, October 22-26, Washington DC.

[3] Zanstra P, Rector A, Ceusters W, de Vries Robbé P. Coding systems and classifications in healthcare: the link to the record. International Journal of Medical Informatics 1998; 48: 103-109.

[4] Ceusters W. Language, medical terminologies and structured electronic patient records: how to escape the Bermuda Triangle. In De Moor G, De Clercq E (eds.) Proc MIC, 2000;: 7-14.

[5] Smith B, Ceusters W. Towards industrial-strength philosophy. How analytical ontology can help medical informatics. Interdisciplinary Science Reviews, 2003; 28 (2): 106-111.

[6] Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C (2004) Relations in biomedical ontologies. Genome Biology, 2005 2005, 6 (5), R46.

[7] Ceusters W, Smith B. Tracking referents in electronic healthcare records. Forthcoming in Proc MIE 2005.

[8] Ceusters W, Smith B. Strategies for Referent Tracking in Electronic Health Records. Journal of Biomedical Informatics. In press. (draft, presented during the IMIA WG6 workshop Ontology and Biomedical Informatics, Rome, Italy, April 29 - Mai 1, 2005)

[9] Powell T, Srinivasan S, Nelson SJ, Hole WT, Roth L, Olenichev V. Tracking meaning over time in the UMLS Metathesaurus. Proc AMIA Symp. 2002;:622-6.

[10] Smith B. Beyond concepts, or: Ontology as reality representation. In: Varzi AC, Vieu L (eds.), Formal Ontology and Information Systems. Proc Third International Conference (FOIS 2004), Amsterdam: IOS Press, 2004;; 73-84.