The VERICLIG Project: Extraction of Computer Interpretable Guidelines via Syntactic and Semantic Annotation

Camilo Thorne, Marco Montali, Diego Calvanese
KRDB Research Centre for Knowledge and Data
Piazza Domenicani 3, Bolzano, Italy
{cthorne,calvanese,montali}@inf.unibz.it

Elena Cardillo, Claudio Eccher
Fondazione Bruno Kessler
Via Sommarive 18, Povo, Italy
{cardillo,eccher}@fbk.eu

Abstract

We consider the problem of extracting formal process representations of the therapies defined by clinical guidelines, viz., *computer interpretable guidelines* (CIGs), based on UMLS and semantic and syntactic annotation. CIGs enable the application of formal methods (such as model checking, verification, conformance assessment) to the clinical domain. We argue that, while minimally structured, correspondences among clinical guideline syntax and discourse relations and clinical process constructs should however be exploited to successfully extract CIGs. We review work on current clinical syntactic and semantic annotation, pinpointing their limitations, and discuss a CIG extraction methodology based on recent efforts on business process modelling notation (BPMN) model extraction from natural language text.

1 Problem Description

Clinical guidelines are evidence-based documents compiling the best practices for the treatment of an illness or medical condition (e.g., lung cancer, flu or diabetes): they are regarded, following Shahar et al. (2004), as a major tool in improving the quality of medical care. More concretely, *they describe or define* the "ideal" (most successful) *care plans or therapies* healthcare professionals should follow when treating an "ideal" (i.e., average) patient for a given illness. Being general, guidelines need to be modified or instantiated relatively to available resources by health institutions, patients or doctors into protocols, and implemented thereafter into clinical workflows or *careflows* within clinical information systems. An important intermediate step for the synthesis of protocols and careflows from guidelines are *computer interpretable guidelines* (CIGs), viz., formal representations of the main control flow features of the described treatment and of its process or plan structure. CIGs can be exploited in a plethora of ways by clinical decision support systems to provide execution support and recommendations to the involved practitioners, guide the refinement into executable clinical protocols and careflows, and check for conformance and compliance.

Clinical document processing, and in particular the authoring of CIGs, protocols and careflows, is however a very costly and error prone task as it involves many layers of manual processing and annotation by experts. This explosion in costs, as pinpointed by Goth (2012), raises the need to develop biomedical NLP techniques, specifically: **(1)** clinical information extraction (IE) techniques and **(2)** automated CIG extraction methodologies.

The VERICLIG project[1], a joint project involving the KRDB Research Centre for Knowledge and Data (Faculty of Computer Science, Free-University of Bozen-Bolzano) and the eHealth group from the Fondazione Bruno Kessler (Trento), intends to address the research problem **(2)** by adopting a computational semantics approach that aims at extracting CIGs from textual clinical guidelines. Our objective is to extract the main control-flow structures emerging from the textual description of guidelines in order to explore, in a second step, the possibility to express them using well-known representation languages.

[1] http://www.inf.unibz.it/~cathorne/vericlig

1.5.1.2 Emphasise advice on healthy balanced eating that is applicable to the general population when providing advice to people with type 2 diabetes.
1.5.1.3 Continue with metformin if blood glucose control remains inadequate and another oral glucose-lowering medication is added.

Figure 1: An excerpt from the NICE diabetes-2 clinical guideline[2]. Each line describes atomic treatments that combine together into a complex therapy.

One such language is the business processing modeling notation (BPMN) standard (see Ko et al. (2009)). Process specification and representation languages allow to leverage on formal methods (verification, model checking) as in Hommenrsom et al. (2008), which are useful for reasoning about the extracted CIGs and relate them with the corresponding executed clinical process. To realize our objective, we build on the work on clinical semantic and syntactic annotation mentioned above as well as on recent efforts on BPMN model extraction by Friederich et al. (2011).

2 Clinical Guidelines and Processes

Clinical guidelines such as, for instance, guidelines related to chronic diseases such as diabetes, allergies or lactose intolerance, are minimally structured documents. They possess however some crucial features: **(1)** they describe a *process*, generically intended as a set of coordinated activities, structured over time, to jointly reach a certain goal, and **(2)** the structure of the process they describe is significantly reflected by English *syntax* and *vocabulary*.

Processes. There are several ways to formally characterize processes, but little consensus as to which is the most appropriate for therapies. Thus, we do not intend at this stage to commit ourselves in the VERICLIG project to a particular formalism, but intend rather to focus on the main features such formalisms share, and in particular on their most basic, common constructs. For convenience, we use the terminology coming from the BPMN standard. In BPMN a process is a complex object constituted by the following basic components:

(*i*) *activities* (e.g., providing advice, controlling blood glucose levels), representing units of execution in the process;
(*ii*) participants, viz., the *actors* (e.g., doctors, nurses, patients), represented using pools, which are independent, autonomous points of execution, and possibly lanes, detailing participants belonging to the same pool;
(*iii*) *artifacts* or *resources* (e.g., metmorfin) used or consumed by activities;
(*iv*) *control flows and gates* (e.g., "if…then…else" control structures) that specify the acceptable orderings among activities inside a pool;
(*v*) *message flows*, representing information exchange between activities and participants belonging to different pools.

Process-evoking Categories. In English, *content words* provide the vocabulary of the domain, denoting the objects, sets and (non-logical) relations that hold therein; their meaning (denotation) is static. On the other hand, *function words* denote the logical constraints, relationships and operations holding over such sets and relations. This distinction holds also to some degree (as allowed by their inherent ambiguity) in clinical domain documents, giving way to *process-evoking word categories* (and constituents).

Figure 1 provides an excerpt taken from a diabetes guideline. In it, activities, actors and artifacts/resources (i.e., static information) are denoted by content words. Activities are denoted often by transitive, intransitive or ditransitive verbs, viz., **VB**s[3] and **VBZ**s, participles (**VBN**s), gerunds (**VBG**s),

[2] `http://www.nice.org.uk/nicemedia/pdf/CG87NICEGuideline.pdf`
[3] In what follows we refer to Penn Treebank word category and syntactic constituent tags, see Marcus et al. (1993).

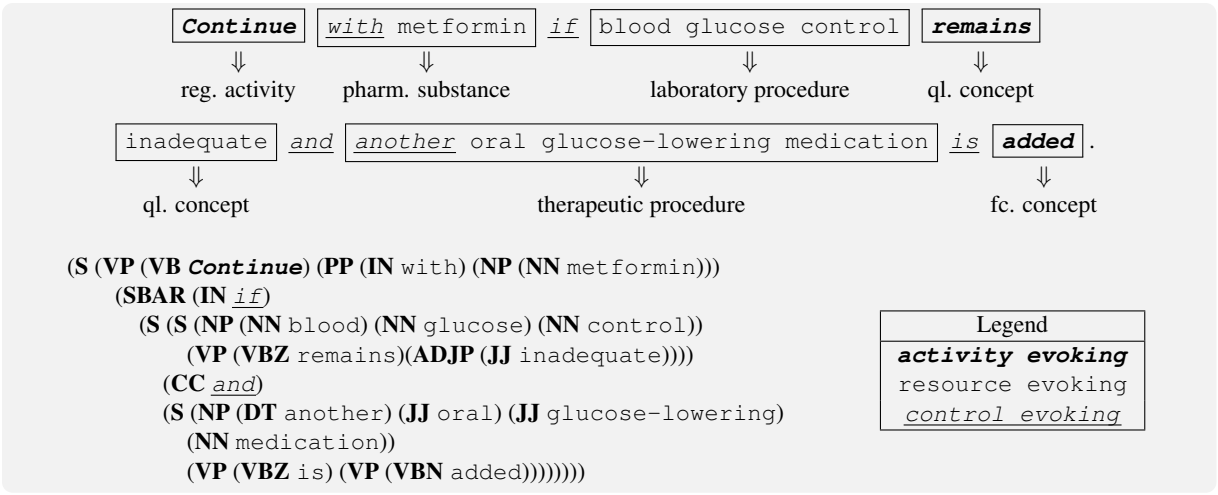

Figure 2: **Top:** MetaMap UMLS (automated) annotation of item 1.5.1.3 from Figure 1. Word highlighting is ours. Entity segmentation (boxes) and annotations are MetaMap's. **Bottom left:** Parse tree obtained with the Stanford parser. Word highlighting is ours.

etc, while actors and resources are denoted by the **NP** complements of such verbs. Control flows (i.e., dynamic, temporal information) are, on the other hand, denoted by function words, i.e., by **(a)** connectives introducing subordinated or coordinated phrases (e.g., **IN**s such as "if", in Figure 1), and **(b)** temporal adverbs (e.g., "following", in Figure 1) or prepositions (e.g., "after", in Figure 1). Such connectives and adverbs are called in NLP literature *discourse relations* since they are used to combine together phrases (noun and verb phrases) and sentences (whether main or subordinated). The correspondences among syntax, discourse relations and clinical process constructs should be exploited to successfully extract clinical processes.

3 Extraction of CIGs: Open Challenges

The main challenge in CIG extraction consists in how to combine semantic annotation techniques, focusing on content words (i.e., on entities and events), with syntactic annotation techniques capable of understanding the control flow structure conveyed by discourse relations, and information extraction methods dealing with clinical English ambiguity. In this section we provide an overview of the research challenges and of our proposed methodology to tackle them.

Limitations of Current Biomedical Resources. Research in biomedical NLP has yielded significant semantic annotation resources. Above all, the US National Library of Medicine's Unified Medical Language System (UMLS) Metathesaurus[4] and the annotated corpora, SemRep and annotation tools built upon it, MetaMap and SemRel, as described by Aronson and Lang (2010). MetaMap is used to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text. MetaMap uses a knowledge-intensive approach based on symbolic, NLP and computational linguistics techniques. Besides being applied for both IE and data-mining applications, MetaMap is one of the foundations of the US National Library of Medicine Medical Text Indexer (MTI) which is being used for both fully and semi automatic indexing of biomedical literature. Other resources that need to be mentioned are the semantically annotated CLEF corpus by Roberts et al. (2007), and Mayo Clinic's Java API cTAKES (version 2.5), by Savova et al. (2010).

Such resources, however, are still of limited use for the CIG extraction task. Gold-standard annotated guidelines are scarce for training and evaluation, and, if available, are not always in the public domain or might not support all biomedical IE tasks. Table 1 shows the main features of the mentioned biomedical

[4] http://www.nlm.nih.gov/research/umls/

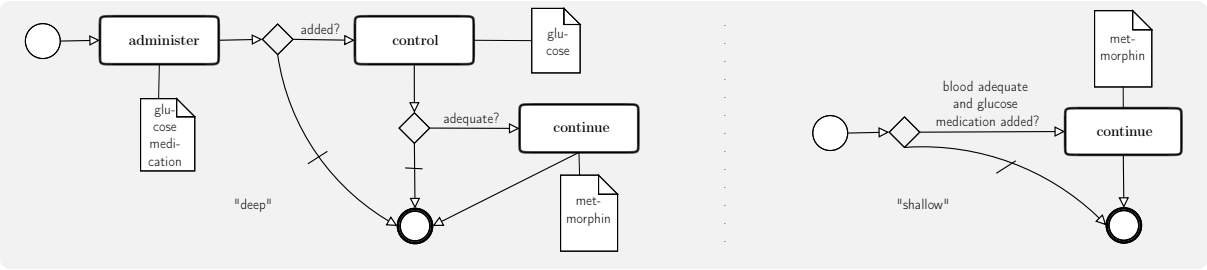

Figure 3: Two possible CIGs (in BPMN notation) of example 1.5.1.3 from Figure 1. Round boxes denote activities, diamonds conditional gates, square boxes resources and edges message flows. Circles represent the end (bold) and beginning of the process (normal), and barred edges "else" conditions.

Resource	ER	TE	RE	ANA	WSD	EV	Type	Open domain
CLEF	✓	✓	✓	✓	×	✓	Annotated corpus	×
UMLS	✓	✓	×	×	✓	×	Lexical resource	× (free for res.)
cTAKES 2.5	✓	✓	✓	✓	×	✓	Java API	✓
SemRep	✓	✓	✓	×	×	✓	Annotated corpus	× (free for res.)

Table 1: Main clinical and biomedical NLP resources and the features/IE tasks they support: entity recognition (ER), term extraction (TE), relation extraction (RE), anaphora resolution (ANA), word sense disambiguation (WSD) and event extraction (EV).

resources.

Crucially, bio-medical/clinical IE systems often overlook process control structure or improperly understand clinical documents. Figure 2 illustrates a SemRel/MetaMap UMLS annotation of example 1.5.1.3 from Figure 1, with its associated phrase structure parse tree. As the reader can see, such tools annotate only the identified "entities", viz., the verbs and **NP**s, overlooking process structure as conveyed by discourse relations. Alternative IE techniques applied to guidelines such as Kaiser et al. (2005) make extensive use of such resources. On the other hand, syntactic parsing, while necessary, as it can identify many discourse relations (e.g., the (**IN** *if*) constituent in Figure 2) and many of their arguments, is not sufficient, due to its limited domain knowledge: it provides little information regarding reified activities (e.g., the UMLS "procedures" in Figure 2). This may give rise to low precision, recall and accuracy for extraction methodologies based on either resource taken *alone*.

Research Steps. The VERICLIG project seeks to understand whether these limitations can be overcome using techniques coming from the business processing community. Friederich et al. (2011) showed how to mine control flow semantics from parse trees (phrase structure and typed dependency trees) computed from business policy documents to extract (generic) business processes in BPMN notation with reasonably high accuracy ($> 70\%$). We would like to adapt their general techniques to the clinical setting, by combining it with biomedical annotations.

For instance, to assign to example 1.5.1.3 the "deep" CIG representation (in BPMN notation) from Figure 3, which we believe accurately captures its semantics, and not the "shallow" one, we need to *combine* the output of syntactic and semantic annotation techniques. SemRep/MetaMap cannot extract process structure, but knows that "medication" and "control" (**NN**s) denote activities and not resources. Parsing, on the other hand, allows us to infer an "if…then…else" control structure, giving rise to the "shallow" process. However, by combining both sources of knowledge, we can see that the subordinated conditional phrase can be broken into a *sequence* of nested "if…then…else" structures. In addition to this, we need to provide support for ambiguity, temporal relations and anaphora resolution, as anaphoric dependencies and temporal relations are needed to build complex models that interconnect the process fragments extracted from each of the guideline's lines. As the reader may infer from Figure 3, process (and hence CIG) components temporally relate to each other, a feature spatially represented in BPMN's graphical notation; thus, we also need to extract such relations. These considerations give rise to the

following CIG extraction methodology:

(*i*) combine annotation resources to extract CIG resources, actors and activities,
(*ii*) analyze syntactic/dependency structure to extract CIG control structure, and
(*iii*) resolve ambiguities and co-references, and infer temporal relations to build a complex CIG.

The work by Friederich et al. (2011) has the advantage, moreover, of proposing alternative methods for measuring, e.g., extraction accuracy, less dependent on the availability of Gold corpora, by comparing the similarity of the extracted model with that of the workflow implemented in business information systems. As both workflows and process representations or models (in, e.g., BPMN notation) are embeddable in graphs, an appropriate evaluation metric is *graph edit distance*. Given the availability of careflows in clinical information systems, this evaluation strategy should be applicable to our case.

In the near future, we intend to implement a baseline system to evaluate our proposed methodology along the lines discussed above. We are currently collecting a corpus of guidelines related to chronic diseases (e.g., diabetes, obesity, food allergy, etc.), and collaborating with local health institutions (the Merano hospital in Merano, Italy) to acquire the required careflows/clinical workflows for CIG extraction evaluation. To further refine CIG extraction, we will also consider applying formal methods (e.g., temporal logic reasoning) to prune logically inconsistent CIGs and/or their components.

References

Aronson, A. R. and F.-M. Lang (2010). And overview of MetaMap: Historical perspective and recent advances. *J. of the American Medical Informatics Association 17*(3), 229–236.

Friederich, F., J. Mendling, and F. Puhlmann (2011). Process model generation from natural language text. In *Proc. of the 23rd Int. Conf. on Adv. Inf. Sys. Eng. (CAiSE 2011)*.

Goth, G. (2012). Analyzing medical data. *Comm. of the ACM 55*(6), 13–15.

Hommenrsom, A., P. Groot, M. Balser, and P. Lucas (2008). Formal methods for verification of clinical practice guidelines. In A. T. T. et al. (Ed.), *Computer-based Medical Guidelines and Protocols: A Primer and Current Trends*, Chapter 4, pp. 63–80. IOS Press.

Kaiser, K., C. Akkaya, and S. Miksch (2005). Gaining process information from cinical practice guidelines using information extraction. In *Proc. of the 10th Int. Conf. on Art. Int. on Med. (AIME 2005)*.

Ko, R. K., S. S. Lee, and E. W. Lee (2009). Business process mangament (BPM) standards: A survey. *Business Process Management Journal 15*(5), 744–791.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics 19*(2), 313–330.

Roberts, A., R. Gaizaskas, M. Hepple, N. Davis, G. Demetriou, Y. Guo, J. Kola, I. Roberts, A. Setzer, A. Tapuria, and B. Wheeldin (2007). The CLEF corpus: Semantic annotation of a clinical text. In *Proc. of the AMIA 2007 Ann. Symp.*

Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute (2010). Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *J. of the American Medical Informatics Association 17*(5), 507–513.

Shahar, Y., O. Young, E. Shalom, M. Glaperin, A. Mayafitt, R. Moskovitch, and A. Hessing (2004). A framework for a distributed, hybrid, multiple-ontology clinical-guideline library and automated guideline-support tools. *J. of Biomedical Informatics 37*(5), 325–344.