

Ontology-driven extraction of linguistic patterns for modelling clinical guidelines

Radu Serban¹, Annette ten Teije¹ Frank van Harmelen¹, Mar Marcos², and
Cristina Polo-Conde²

¹ AI Department, Vrije Universiteit, The
Netherlands, {serbanr, annette, frankh}@few.vu.nl

² Departament d' Enginyeria i Ciència dels Computadors, Universitat Jaume
I, Castellón, Spain, Mar.Marcos@icc.uji.es, Cristina.Polo@sg.uji.es

Abstract. Evidence-based clinical guidelines require frequent updates due to research and technology advances. The quality of guideline updates can be improved if the knowledge underlying the guideline text is explicitly modelled using the so-called **guideline patterns (GPs)**, mappings between a text fragment and a formal representation of its corresponding medical knowledge.

Ontology-driven extraction of linguistic patterns is a method to automatically reconstruct the control knowledge captured in guidelines, which facilitates a more effective modelling and authoring of clinical guidelines. We illustrate by examples the use of a method for generating and searching for linguistic guideline patterns in the text of a guideline for treatment of breast cancer, and provide a general evaluation of usefulness of these patterns in the modelling of the guideline analyzed.

1 Introduction

Authoring and maintenance of medical guidelines is recognized as an important factor for improving the quality of healthcare, but also a costly process.

Medical guidelines change frequently due to research and technology improvements, but only parts of the guideline need updates. Guideline formalization makes explicit a modular organization of medical knowledge and produces an executable model from the guideline recommendations, therefore facilitating a more effective update of the guideline knowledge and verification of guideline properties. To avoid repeating guideline formalization from scratch each time a guideline is updated, recent research ([3, 13]) suggests to split the formalization into several steps, isolating procedural, medical, organizational knowledge and defining the so-called **guideline patterns (GPs)**, which represent mappings between text fragments and a more formal representation of its underlying knowledge.

A few linguistic constructs are frequently recurring in the text of clinical guidelines, regardless of the domain addressed by the guideline. For instance,

This work has been supported by the European Commission's IST program, under contract number IST-FP6-508794 Protocure-II.

conclusions and recommendations typically have a modular structure, easy to recognize and useful in modelling the guideline.

If linguistic regularities such as these:
In the event of [MedContext], the treatment of choice is [Treatment]. or
In the event of [MedContext], [Treatment] is recommended.
 can be given a formal representation, it seems natural to define knowledge templates that are instantiated by these statements, which can be reused when making new guidelines or changing a particular type of knowledge. These so-called linguistic pattern templates help us in establishing a set of modular components for modelling guidelines in the form of: (1) a shared vocabulary and (2) a language to describe linguistic regularities conveying a specific type of knowledge. This mapping between text and the knowledge underlying it makes validation of this knowledge straightforward and eases the modelling task. Authoring and updating of guidelines can also benefit from these modular components, as only the parts concerned with a changing piece of knowledge need updated.

In this paper we focus on knowledge templates that describe control (procedural) knowledge, and investigate their role in improving modularization and formalization of clinical guidelines. We propose a method that uses linguistic regularities in the text of a guideline, and an ontology of the medical domain, to generate a list of linguistic templates, which is explained in section 2 and summarized in figure 1. In section 3 we discuss our algorithm for searching instances

Algorithm 1.1
GUIDELINE-FORMALIZATION(TF,PT)
 ▷ *TF*:text fragment; *PT*:set of patt.templates

1. build an ontology from a corpus of guideline texts;
2. semantically tag the guideline text using the ontology: replace terms in the text with their corresponding ontological categories;
3. generate control templates using the ontological categories identified;
4. select a set of core templates, by eliminating templates covered by or made of other templates;
5. establish a formal translation for the core templates;
6. find instances of core templates in the guideline;
7. translate pattern instances into their formal equivalent using the translation pattern of their corresponding template;
8. select the linguistic templates instantiated in more than one guideline text, as guideline building blocks.

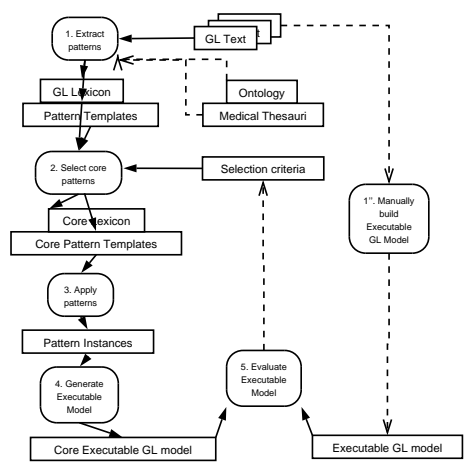


Fig. 2. Steps for extracting and evaluating linguistic patterns

Fig. 1. Guideline formalization using patterns

of linguistic patterns and their use in the guideline formalization. In section 4 we evaluate the effectiveness of pattern detection in generating an executable model of a breast-cancer guideline. Section 5 presents related work and section 6 summarizes the paper contribution, emphasizing the benefits of using linguistic patterns as support for guideline formalization. Figure 2 depicts the steps

reverse engineering of the domain addressed by the guideline, since a part of the formal representation of the guideline is represented by these relations.

The example in figure 4 illustrates how pattern templates can be extracted from the text of a recommendation taken from the 2002 **CBO guideline for treatment of breast cancer** ([5]). If we replace instances present in the recommendation (1) with their categories in the ontology, we obtain a skeletal representation of the sentence. This intermediate representation of a pattern template contains concepts from an ontology and terms from a non-medical lexicon (2). If we apply the categorization rules in the ontology (which contains relations such the ones depicted in figure 3), to represent the sentence skeleton at a higher level of abstraction, the recommendation is rewritten as expression (3). Finally, if we ignore the linking words (of the lexicon) and consider only the categories present in the ontology, we obtain a more compact template of the recommendation, as depicted in expression (4).

- ↓ refined_as
- (1) *Recommendation*: {Patients with} [disease] {should receive} [treatment] {with} [med_goal].
- ↓ refined_as
- (2) {Recommendation}: [Target_group] [recommendation_op] {receive} [treatment] {with} [med_goal].
- ↓ refined_as
- (3) {Recommendation}: [med_context] [recommendation_op] [complex_treatment].

Fig. 4. Abstraction steps for extracting a pattern template

ommendation text (1) with their categories in the ontology, we obtain a skeletal representation of the sentence. This intermediate representation of a pattern template contains concepts from an ontology and terms from a non-medical lexicon (2). If we apply the categorization rules in the ontology (which contains relations such the ones depicted in figure 3), to represent the sentence skeleton at a higher level of abstraction, the recommendation is rewritten as expression (3). Finally, if we ignore the linking words (of the lexicon) and consider only the categories present in the ontology, we obtain a more compact template of the recommendation, as depicted in expression (4).

The recommendation contains an instance of *med_context* ("Patients with locoregionally advanced breast cancer") followed by a *recommendation_op* ("should") and an instance of *med_action* ("receive multidisciplinary treatment with curative intent"); the latter can be further refined as a sequence of: *treatment* ("multidisciplinary treatment") followed by *med_goal* ("with curative intent"). The advantage of having such a conceptual sketch of the linguistic construct "med_recommendation" is that the template of any recommendation will include one of the following ordered lists of medical categories, obtained by refining parts of the linguistic component:

(*med_context, recommendation_op, med_action*)
(*target_group, recommendation_op, treatment, med_goal*), and so on.

The goal of finding linguistic templates in the text requires us to find n-grams with elements belonging to either a medical category, such as *target_group* or *med_goal*, or to a lexical category such as *ctx_op*, *recommendation_op*, which links medical terms. Disambiguation of some of the terms is required, nonetheless the use of a terminology system when authoring the guidelines would reduce the importance of this task. By filtering the detected n-grams using the relevant semantic relations provided by the ontology, a grammar for defining linguistic pattern templates can be derived. Even though pattern templates can be gen-

erated and instantiated automatically using this method, producing meaningful linguistic pattern templates cannot be fully automated.

3 Detection of pattern instances in the guideline text

For identifying instantiations of existing pattern templates in the guideline text we use our custom built ontology of the medical domain, summarized in section 2. Figure 3 contains a few examples of concepts from this ontology:

1. **medical specific categories:** disease, medication, body_part, med_effect, med_action;
2. **operator categories** - lexical terms corresponding to semantic relations between medical categories in the ontology: relational operators (assoc_rel_op, temp_rel_op, causal_rel_op) or action operators (decomp_op, act_op)

We built an application that generates templates as sequences of medical concepts that are connected using control relations in the ontology, for instance: *template([med_action, effect_op, med_effect]) covers*

ontology_fragment(MedAction produces MedEffect)

We define a set of control relations relevant for the operational model of the guideline: causal relationships between actions, ordering and decomposition of actions, correlations condition-action, action-intention, action-effect, etc. We transform them into pattern templates and look up their instances in the guideline.

Implementation. The guideline text is split into sentences and further into word-level chunks. A *guideline chunk* is a pair $\langle TF, Ann \rangle$, where *TF* represents a text fragment potentially relevant for the pattern detection, and *Ann* is a list of semantic annotations for *TF*. Initially, the chunk is set at the word level, but during semantic annotation, it can group together several words and even sentences, depending on the level of granularity at which patterns are recognized.

A pattern template is the abstraction of a text fragment as a list of concepts from two sources: a medical ontology and a non-medical lexicon containing frequent link words. We define patterns at different levels of granularity: (1) patterns at word-level are in fact semantically tagged medical terms in the guideline text; (2) pattern at sentence level define concepts from different semantic categories which correspond to well-defined formal constructs.

For instance, the sentence "Treatment1 consists of axillary surgery followed by radiotherapy." consists of two basic (core) patterns: *action₁ consists_of action₂, action₃*, a hierarchical decomposition, and *action₂ followed_by action₃*, an action sequencing.

Initially, the sentence is split into word-level chunks, and the list of annotations of each word contains only the relative position of the term in the guideline text. At each processing of the sentence, in search for patterns, the annotations can be expanded as follows: when the term of the chunk is an instance of a medical term, its semantic categories are added to the annotation list; when a pattern is recognized, of which the chunk can be a part, it is added as annotation of that

chunk, etc. When medical terms are recognized in the sentence, several chunks corresponding to the component words are merged into one chunk, together with their annotations. In the next step the analysis focuses on sentence-level chunks. Sentence-level chunks are sequences of word-level chunks, annotated with possibly overlapping pattern instances found in that sentence.

Results of pattern detection. The parameters of pattern detection are: (1) the medical ontology; and (2) a set of target pattern templates sought in the text. After applying the algorithm described above for the reference guideline ([5]), and reviewing the instances found, the most frequent operational patterns were: $p_{1.1}(A : med_action, \{following\} : seq_act_op, B : med_action)$ and $p_{1.2}(A : med_action, \{after\} : act_op, B : med_action)$.

They are subclasses of a more abstract pattern - sequence of two medical actions, denoted: $p1(med_action, seq_act_op, med_action)$. In figure 5 we have depicted the template $p1$ corresponding to three pattern instances $i1, i2, i3$. Pattern $p1$ says that a frequent template consists of an ordered list of slots, of which the first and the third one can be filled with instances of medical actions, and the middle one can be filled with any instance of an action operator, describing relations between actions. For instance, in chapter 3, 134 out of 179 sentences were deemed relevant for analysis, and 226 such pattern instances (including overlappings) were identified.

```
i1(axillary_surgery, following, excision))
i2(biopsy, following, excision))
i3(breast_reconstruction, following, mastectomy))
  ↓ instance_of
p1(med_action, act_op, med_action)
```

Fig. 5. *Pattern template extracted from several instances*

By grouping together pattern instances that instantiate same templates or share common words, the most frequent linguistic constructs can be retrieved and reused in guideline authoring and formalization. For instantiations of control patterns, an equivalent executable representation can be generated automatically, based on the translation of the underlying pattern template into actions.

Selection of core patterns. The process of pattern detection produces a list of pattern templates and a core lexicon of link words that connect medical terms in the pattern instances detected. For the guideline analyzed ([5]), the lexicon contains link words such as:

```
conditional_op : if, in_the_case_of, in_the_event_of
effect_op : results_in, improves, is_expected_to
sequential_op : after, following, followed_by, before, initially
causal_op : since, because, due_to
recommendation_op : should, is_recommended, advisable_to
```

The list contains relational operators grouped according to the type of semantic relation they describe: ordering of actions, quantification of action effects, etc.

Relations between linguistic templates. After the instances of medically-relevant pattern templates have been looked up in the guideline text, we choose as basic pattern templates those which have the highest support and are more abstract than other templates.

A semantic annotation $SemAnn : T_{GL} \rightarrow Cat$ of the guideline text T_{GL} produces a list of semantic categories from the set Cat . Medical background knowledge expected in the guideline is represented as a set of facts BK about elements in Cat . A schema is a collection of primitive items in Cat connected by relations between items or sets of items. The set of all schemas produced by Cat is denoted S_{Cat} . A schema $S \in S_{Cat}$ is called maximal if it is not a subschema of any other schema $S_1 \in S_{Cat}$. Pattern templates with a high level of abstraction represent maximal schemas. For selecting the core templates, we define relations between linguistic patterns $PT_1 = [C_{11}, C_{12}, \dots, C_{1n}]$ and $PT_2 = [C_{21}, C_{22}, \dots, C_{2n}]$, using the hierarchical relations in the ontology:

is-more-specific(PT_1, PT_2) iff for all $i = \overline{1, n}$: $is_a(C_{1i}, C_{2i})$;
contains(PT_1, PT_2) iff $\{C_{21}, C_{22}, \dots, C_{2n}\} \subset \{C_{11}, C_{12}, \dots, C_{1n}\}$.

The list of core pattern templates with high coverage among the instances identified is depicted in table 1, with frequencies for three guideline chapters used in the evaluation in section 4.

Template	Translation	Ch.2-4
Association action-goal : $[med_action, assoc_rel_op, disorder]$ $[surgery, to_reduce, tumour_load]$	action-goal	10
Action decomposition : $[med_action, decomp_op, med_action, med_action]$ $[current_treatment, consists_of, surgery, radiotherapy]$	decomp.	3
Association condition-action : $[med_context, med_action]$ $[multidisciplinary_treatment, chemotherapy]$	if-then	12
Action sequencing : $[med_action, act_op, med_action]$ $[radiotherapy, following, neoadjuvant_chemotherapy]$	sequencing	29
Associations action-effect : $[disorder, temp_rel_op, med_action]$ $[tumour_recurrence, following, radiotherapy]$	action-effect	2
Preference for actions : $[treatment, assoc_rel_op, med_action]$ $[treatment_of_choice, is, neoadjuvant_chemotherapy]$	preferences	19

Table 1. Coverage of core pattern templates in the chapters analyzed

4 Evaluating the use of patterns in guideline formalization

Guideline formalization is a transformation that takes as input a guideline GL and a set of formalization rules RF, and produces an executable representation E of the procedural part of the guideline. Formalization involves the following steps: [1.] select a set of control relations relevant for the target model, then generate templates corresponding to these relations; [2.] detect instances of the control templates in the guideline text; [3.] transform these instances into their formal equivalent. The text-driven approach to formalization consists of deriving

a set of constraints RF by reverse-engineering, using a domain-specific lexicon, of the mappings between text fragments and medical knowledge, and using the representation of that knowledge in the guideline representation language to obtain E. To evaluate how close two executable models are, in this paper we make a simplifying assumption: an executable representation of a guideline consists of the actions and the control relations referenced in the guideline.

Evaluation results. We have compared the results of modelling chapters 2, 3 and 4 of the **CBO guideline for treatment of breast cancer** ([5]) in the medical language MHB [14], using two methods: one which generates a guideline model from pattern instances found automatically as described in this paper, and one which employs a human knowledge engineer (KE) to build the model manually. To estimate the usefulness of applying patterns in guideline formalization, the executable model produced using the linguistic patterns identified automatically is evaluated against and expected to be aligned with the "golden standard" model produced by the human modeller.

We used only instances of templates denoting control relations: action sequencing and decomposition, which were deemed relevant for a medical executable model. To assess if these patterns are suitable to be used for knowledge acquisition in the beginning of guideline formalization, we evaluated whether it is possible to build a coherent fragment of an executable MHB model from the pattern instances detected. The evaluation consisted of: (1) a rough comparison (quantitative) of the amount of knowledge (automatically) identified by using patterns with respect to the knowledge modelled by (manual) knowledge acquisition; for this, we compared the amount of sentences in which the pattern search application has found patterns with respect to the sentences modelled by the KE as procedural knowledge. (2) an analysis (qualitative) of the utility of the pattern instances identified in specific fragments of the guideline; we studied whether a significant piece of a medical executable model can be directly obtained from the pattern instances. This gives an indication of the potential of the pattern detection process for knowledge acquisition.

	(autom.) processed sentences	(manual.) modelled sentences	modelled sentences processed	modelled sentences with patterns
chapter 2	130	41	30 (73%)	8 (19.5%)
chapter 3	134	20	16 (80%)	7 (35%)
chapter 4	91	25	18 (72%)	7 (28%)

Table 2. *Evaluation: linguistic patterns vs. manual annotation in modelling guidelines*

We have evaluated the coverage of the detection process with respect to the procedural parts modelled by the KE by calculating the percentage of sentences where patterns were detected. Table 2 shows the numbers obtained for the different chapters modelled. Column 1 shows the number of sentences processed by the application and considered relevant for the guideline topic, using a keyword list as criteria for relevance. Columns 2 and 3 give respectively the number of sen-

tences actually modelled by the KE (i.e. the sentences considered relevant from the KE's viewpoint) and, among them, the amount of sentences processed by the application (both the number and the percentage with respect to the modelled sentences). Finally, the last column shows the amount of sentences modelled by the KE and also processed by the application where some patterns have been found. The amount of sentences considered relevant by the application exceeds the modelled knowledge, but covers it to a significant extent, between 70% and 80%. The relatively low coverage of the executable model is explained by the low granularity of the automatically detected patterns, and the absence of some semantic relations from the ontology. Other obstacles in automatic detection were the use of tables and references to non-medical actions and terms absent from the ontology, that could not be extracted from tables. Better coverage heavily depends on having a complete classification of medical terms, particularly actions. Using a richer thesaurus for generating the skeleton executable model from patterns would prove helpful in supporting formalization.

5 Related Work

Guideline patterns reflect modelling decisions when medical guidelines are transformed into an executable form. The task of extracting structure and semantics from annotated and unannotated text, for supporting querying and natural language understanding, has been addressed by recent research in text and data mining (see, for instance, the MedLEE system and related work [7]). The main trend has been extraction of vocabularies or simple syntactic constructs from untagged text ([6, 10–12]), in some cases guided by the use of a dictionary, thesaurus, or positive examples ([8]). Assigning domain-specific categories to text is done using background knowledge in the form of conceptual graphs ([15, 16]) or simpler mappings between concepts in an ontology and terms in the target domain of the textual description. Statistical and probabilistic models ([6]) were used to increase the performance when ambiguous textual constructions are present. Our work has similarities with concept and relation extraction, but focuses on the use of an ontology to generate, not only validate, pattern candidates for a category of texts with rather strict formatting rules.

6 Conclusions

Searching linguistic patterns is motivated by the need for reusable guideline blocks in guideline formalization and authoring, and by the high overlap between the medical vocabularies used by the oncology guidelines analyzed. The pattern search process is guided by the mappings between medical terms and concepts in a medical ontology, which help us to: 1. extract control knowledge from text, in the form of pattern templates; 2. select a set of core pattern templates, using pattern relationships; 3. identify pattern instances for existing pattern templates. The process takes as input the text of an existing guideline, and an ontology,

and attempts to reverse engineer the recurring linguistic pattern templates containing those terms from the ontology that were used to produce the text. Linguistic patterns are basic building blocks from which semantically-rich fragments can be built, facilitating modularization, validation and reuse of the background knowledge covered by guidelines. The use of patterns produces a lexicon and a skeleton of the formal model covered by the procedural part of the guideline, automatically. The method proposed can be extended to non-procedural knowledge, therefore authoring and formalization of clinical guidelines can benefit from the use of the ontology-driven approach to obtaining linguistic patterns.

References

1. Mesh (Medical Subject Headings). URL: <http://www.nlm.nih.gov/mesh/meshhome.html>.
2. National cancer institute ontology. URL: <http://www.mindswap.org/2003/CancerOntology/>.
3. Protocure 2 project. URL: www.protocure.org.
4. Unified medical language system. URL: <http://www.nlm.nih.gov/research/umls/>.
5. CBO. *Guideline for the Treatment of Breast Carcinoma*. 2002. PMID: 12474555.
6. Paynter G.W. Witten I.H. Gutwin C. Frank, E. and C. Nevill-Manning. Domain-specific keyphrase extraction. In *Procs. Int. Joint Conf. on AI*, pages pp.668–673. Morgan Kaufmann Publishers, San Francisco, CA, 1999.
7. C. Friedman and G. Hripcsak. Evaluating natural language processors in the clinical domain. In *Procs. Conf. on Natural Language and Medical Concept Representation*, pages 41–52. IMIA WG6, 1997.
8. Scott B. Huffman. Learning information extraction patterns from examples. In *Learning for Natural Language Processing*, pages 246–260, 1995.
9. Y.; Johnson P. Miksch, S.; Shahar. Asbru: A task-specific, intention-based, and time-oriented language for representing skeletal plans. In *Procs. 7th W-shop on Knowledge Engineering: Methods and Languages (KEMML-97)*, 1997.
10. Antonion Moreno and Chantal Perez. From text to ontology: Extraction and representation of conceptual information. In *Procs. Conference on TIA*, May 2001.
11. E. Riloff. Automatically generating extraction patterns from untagged text. In *Procs. 13th Nat. Conf. on AI (AAAI-96)*, pages 1044–1049, 1996.
12. Ellen Riloff and Jay Shoen. Automatically acquiring conceptual patterns without an automated corpus. In *Procs. 3rd W-shop on Very Large Corpora*, pages 148–161, New Jersey, 1995. Assoc. for Computational Linguistics.
13. Svatek V. Ruzicka M. Mark-up based analysis of narrative guidelines with the stepper tool. In *Procs. Symposium on Computerized Guidelines and Protocols (CGP-04)*. IOS Press, 2004.
14. Andreas Seyfang, Sylvia Miksch, and Peter Votruba. Specification of formats of intermediate, asbru and kiv representations. TR-D2.2a, Protocure-II, June 2004.
15. Ian Witten. Adaptive text mining: Inferring structure from sequences. *J. of Discrete Algorithms*, 2000.
16. Jiwei Zhong, Haiping Zhu, Jianming Li, and Yong Yu. Conceptual graph matching for semantic search. In *ICCS*, pages 92–196, 2002.