Artificial Intelligence in Medicine (2006) **xxx**, xxx—xxx

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Extraction and use of linguistic patterns for modelling medical guidelines[☆]

Radu Serban [a,*], Annette ten Teije [a], Frank van Harmelen [a], Mar Marcos [b], Cristina Polo-Conde [b]

[a] *Artificial Intelligence Department, Vrije Universiteit, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands*
[b] *Department of Computer Engineering and Science Department, Universitat Jaume I, Castellon, Spain*

Received 16 January 2006; received in revised form 26 July 2006; accepted 28 July 2006

KEYWORDS
Knowledge engineering;
Ontologies;
Medical guideline formalization;
Semantic mark-up

Summary

Objective: The quality of knowledge updates in evidence-based medical guidelines can be improved and the effort spent for updating can be reduced if the knowledge underlying the guideline text is explicitly modelled using the so-called **linguistic guideline patterns**, mappings between a text fragment and a formal representation of its corresponding medical knowledge.
Methods and material: 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 medical guidelines. We illustrate by examples the use of this method for generating and instantiating linguistic patterns in the text of a guideline for treatment of breast cancer, and evaluate the usefulness of these patterns in the modelling of this guideline.
Results: We developed a methodology for extracting and using linguistic patterns in guideline formalization, to aid the human modellers in guideline formalization and reduce the human modelling effort. Using automatic transformation rules for simple linguistic patterns, a good recall (between 72% and 80%) is obtained in selecting the procedural knowledge relevant for the guideline model, even though the precision of the guideline model generated automatically covers only between 20% and 35% of the human-generated guideline model. These results indicate the suitability of our method as a pre-processing step in medical guideline formalization.
Conclusions: Modelling and authoring of medical texts can benefit from our proposed method. As pre-requisites for generating automatically a skeleton of the guideline

☆ This is an extended and revised version of our paper presented in the Conference on Artificial Intelligence in Medicine (AIME 05).

* Corresponding author. Tel.: +31 20 598 7818; fax: +31 20 598 7653.
E-mail addresses: serbanr@few.vu.nl (R. Serban), annette@few.vu.nl (A. ten Teije), frankh@few.vu.nl (F. van Harmelen), Mar.Marcos@icc.uji.es (M. Marcos), Cristina.Polo@sg.uji.es (C. Polo-Conde).

0933-3657/$ — see front matter
doi:10.1016/j.artmed.2006.07.012

model from the procedural part of the guideline text, to aid the human modeller, the medical terminology used by the guideline must have a good overlap with existing medical thesauri and its procedural knowledge must obey linguistic regularities that can be mapped into the control constructs of the target guideline modelling language.

1. Introduction

1.1. Background

Medical guidelines have been recognized as important instruments for improving the quality of health care by reducing the practice variability and containing the costs of care. Due to their frequent pace of change, influenced by research and technology advances and by new clinical trials, their authoring and maintenance is a challenging knowledge engineering problem. This resource-intensive process can be streamlined if the knowledge is structured along those knowledge components which are most likely to change, and the changes can be tracked to the original medical knowledge which underlies each recommendation.

To handle such changes of the guideline text which affect specific types of medical knowledge, guideline formalization has been employed, which produces a so-called formal model (logical or executable representation) in close connection with the recommendations of the guideline. But the formalization process is not yet sufficiently structured to produce modular medical knowledge in a systematic way, which would allow mapping of this knowledge to the guideline document structure, making possible an effective update of the guideline knowledge and the verification of its properties. To improve the guideline formalization process and to avoid repeating it from scratch each time a guideline is updated, recent research [1—4] suggests to split the formalization into several steps, isolating procedural and declarative knowledge and defining the so-called **linguistic guideline patterns**, which represent mappings between text fragments and a more formal representation of its underlying knowledge.

Guideline texts can be seen as collections of clinical argumentations, therefore the types of knowledge and the principles of structuring this knowledge used in general scientific argumentations can be used. Uren, Shlum et al. [5] suggest that three kinds of knowledge support generic domain-related argumentation in scientific literature, including thus medical guidelines: (a) terminological knowledge—the vocabulary used to describe domain concepts and relationships; (b) domain descriptive knowledge—specific knowledge required to solve problems in the domain; this can be causal, qualitative—descriptive or quantitative knowledge; (c) problem solving knowledge—'how-to' knowledge that allows the knowledge to be applied to problem solving in the domain. Terminological knowledge has been so far the most investigated [6], and methods for effective mapping of medical text to existing thesauri are available [7,8]. As Hahn et al. [9] note, processing of medical narratives has to include lexical relations which underlie the knowledge relations between text fragments.

1.2. Objective

Our goal is to facilitate guideline formalization by reducing the effort spent in manual modelling, particularly that of procedural knowledge captured by guidelines in a narrative form. We try to establish patterns for formal translation from text to a medical knowledge representation language, by observing regularities in the text and by mapping them to control structures in the target medical representation language. If a sufficiently high percentage of the narrative text in the guideline conforms to linguistic regularities for which transformations into elements of a guideline representation language can be identified, then the use of these knowledge transformation patterns would greatly reduce the effort spent in modelling the guideline recommendations.

Certain linguistic constructs are frequently recurring in the text of medical guidelines, regardless of the domain addressed by the guideline. For instance, conclusions and recommendations typically have a modular structure, easy to recognize and useful in modelling the guideline, such as these:

In the event of [*MedContext*], *the treatment of choice is* [*Treatment*], *or*
In the event of [*MedContext*], [*Treatment*] *is recommended*.

If such linguistic regularities 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 templates, or so-called **linguistic patterns** help us in establishing a set of modular components for modelling guidelines in the form of: (1) a controlled vocabulary of lexical markers and (2) a language to describe linguistic regularities conveying a specific type of knowledge. This mapping between the text and its underlying semantic interpretation makes validation of medical 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 to be updated and the textual representation of a piece of medical knowledge can be generated automatically.

In this paper we investigate the role of knowledge templates describing procedural knowledge in improving modularization and formalization of medical 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. In Section 3 we discuss our algorithm for searching instances 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.

2. Approach

We propose to use linguistic patterns in the formalization of medical guidelines, to reduce the effort spent in modelling of a guideline. This section discusses our method for building a set of linguistic templates which are then applied to support the automatic translation of the guideline text into a guideline modelling language representation. Fig. 1 illustrates the steps we performed to

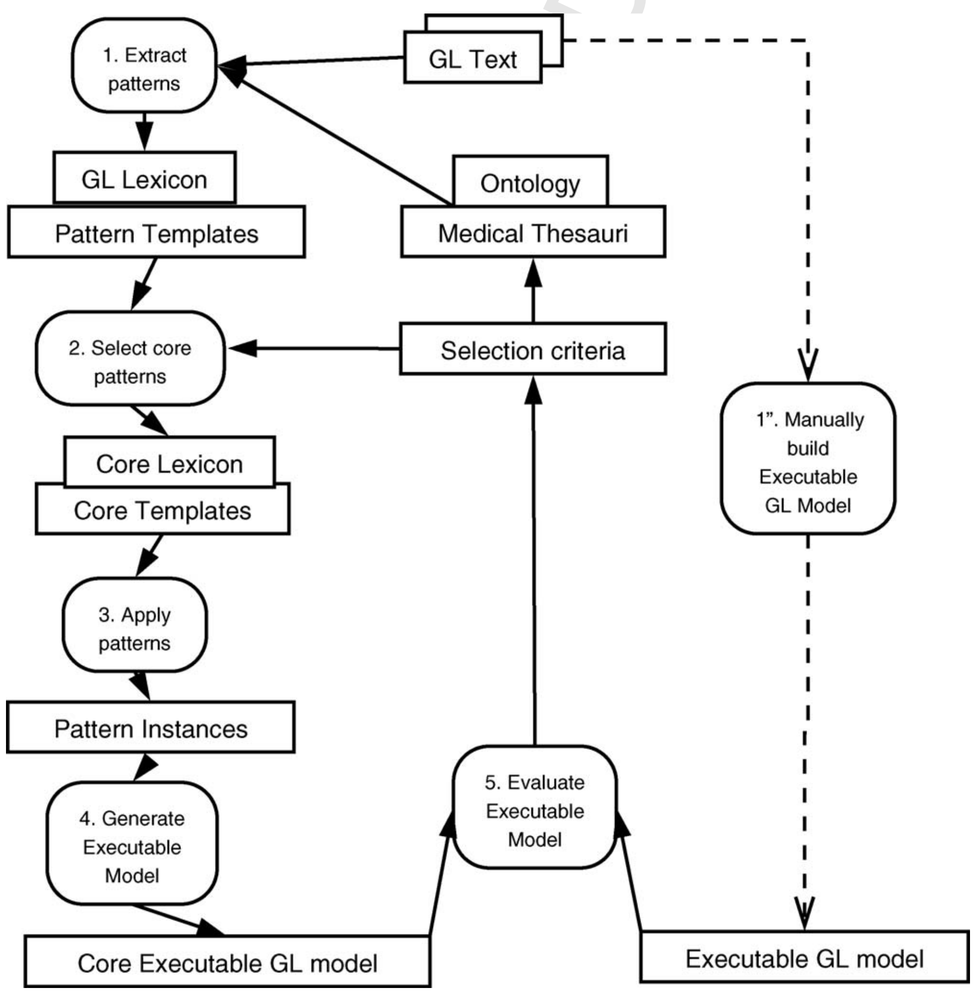

Figure 1 Evaluating the use of linguistic patterns in guideline formalization vs. manually built golden standard (right).

Please cite this article as: Radu Serban et al., Extraction and use of linguistic patterns for modelling medical guidelines, Artificial Intelligence in Medicine (2006), doi:10.1016/j.artmed.2006.07.012

build a set of core linguistic templates and to evaluate their use in guideline formalization. Activities are marked as rounded rectangles, the objects produced by them are shown as plain rectangles. Each activity is discussed in one of the subsequent sections.

We propose a methodology of guideline formalization using linguistic patterns, as illustrated by the procedure depicted in Fig. 2. Steps 1, 2 and 3 correspond to the activity 1 (extraction of patterns) of Fig. 1, step 4 corresponds to activity 2 of Fig. 1, and steps 5 and 6 correspond to activity 4. This work is inspired by recent proposals to use semantic mark-up for processing narrative procedural fragments [3,4] and to identify reusable textual components [10]. The procedure *ExtractLinguisticTemplates* is described in Section 3 and *FormalizationUsingLinguisticTemplates* is discussed in Section 4.

3. Extraction and instantiation of linguistic templates

3.1. Normalizing and semantic tagging of the guideline using a domain model

Our method uses background knowledge about the medical and guideline representation domain which determines how linguistic regularities occurring in the medical text are transformed into corresponding fragments in a guideline representation language. We have chosen ASBRU [12] and Multi-Headed Bridge (MHB) language [13] from a list of guideline modelling languages [11], but our methodology is applicable to other guideline languages as well. Steps 1 and 2 of the algorithm in Fig. 2 use an ontology of the medical domain, DO, to recognize the most frequently encountered templates. The domain knowledge can be represented in a graphical

Algorithm 1

GUIDELINEFORMALIZATION(IN TF,DO,CO; OUT FR)

Parameters used: TF:guideline text fragment; DO: domain ontology; CO: control ontology; FR:formal representation; PT: set of linguistic templates

I. ExtractLinguisticTemplates(in TF,DO,CO; out PT)

1. normalize TF and look up its most frequently used medical terms in DO;

2. semantically tag TF using DO: map and replace the terms in TF with their corresponding DO ontological categories;

3. generate control templates using the CO relations between DO ontological categories identified:

3.1. based on relations between the DO medical concepts encounteredmost frequently, generate domain knowledge templates;

3.2. based on CO relations, generate templates conforming to the guideline modelling language constructs which contain similar relations or concepts as the domain-specific templates;

4. select a set of core templates, by eliminating templates derived from other ones;

4.1. detect template instances in TF and establish the most frequently instantiated templates;

4.2. refine the instantiated templates by using combination and by establishing relations between them, using DO and CO, then select a set of core templates

4.3. map the elements of the core templates to constructs allowed by the target guideline representation language.

II. FormalizationUsingLinguisticTemplates(in PT,CO,DO; out FR)

5. establish the formal translation of the core templates, then derive a formal translation for the derived patterns, when possible;

6. apply the transformation patterns to template instances identified, to obtain a formal translation of the procedural fragments with linguistic regularities.

Figure 2 Steps for extraction and use of patterns in guideline formalization.

Please cite this article as: Radu Serban et al., Extraction and use of linguistic patterns for modelling medical guidelines, Artificial Intelligence in Medicine (2006), doi:10.1016/j.artmed.2006.07.012

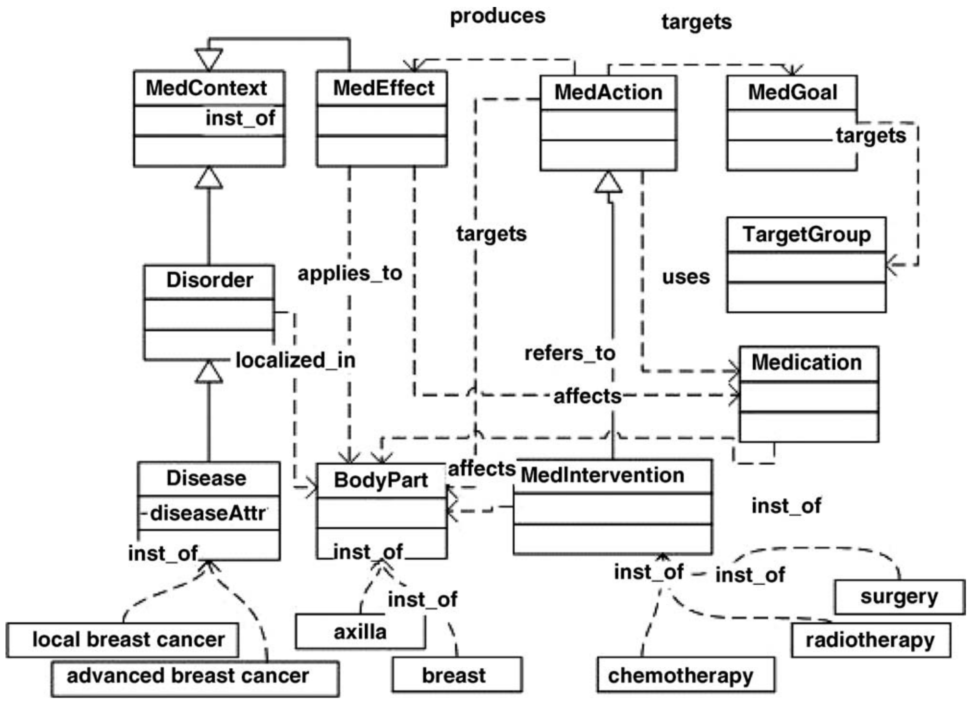

Figure 3 Relations between medical terms and concepts in the medical ontology.

form, as depicted in Fig. 3. This light-weight medical ontology contains "*is_a*" hierarchical relations (subClassOf inclusion) between medical concepts, "*instance_of*" relations between medical concepts and medical terms (concrete medical terms that can be encountered in the guidelines are viewed as labels associated with an "instance of" the specified concept; this is similar to the relation between medical terms to Concept Unique Identifiers (CUIs) of the Unified Medical Language System (UMLS) [18]), and labelled (non-hierarchical) semantic relations between medical concepts. This ontology can be built from existing ontologies (WordNet [14], GALEN [16], SNOMED [15]) or medical thesauri (MeSH [17], UMLS [18], NCIOncology [19] or medical dictionaries [20]). In our case, we built a customized ontology which contains classes from UMLS, and is designed to represent a rather generic description of the medical domain. This ontology is built upon on a subset of the semantic network of UMLS, enriched with medical relations, and is suitable as domain knowledge for a large category of medical texts which require semantic interpretation.

3.2. Generating pattern templates

To capture the procedural aspect of a guideline, we generate control templates using the relations allowed by the ontology of the guideline representation language, such as: action sequencing, decomposition or condition-action, etc.

As part of step 3 in the algorithm of Fig. 2, we perform the following steps. Initially, simple templates are generated, such as: a medical action, followed by a control operator (e.g., sequencing) and an additional medical category (e.g., medical action or goal). Then more specific and complex templates could be generated by (1) adjusting the level of abstraction of the concepts in the pattern, (2) replacing them with specific ones, (3) replacing by operators the semantic relations among several concepts from the ontology, or (4) merging two instances of simpler templates (e.g., if they share a word).

Guideline developers use the implicit knowledge captured by these semantic relations when producing the guideline. Guideline formalization can benefit from 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 Fig. 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**[21]. If we replace the terms present in the recommendation text (row 1) in Fig. 4 with their category tags 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

Please cite this article as: Radu Serban et al., Extraction and use of linguistic patterns for modelling medical guidelines, Artificial Intelligence in Medicine (2006), doi:10.1016/j.artmed.2006.07.012

(1) *Recommendation: Patients with locoregionally advanced breast cancer should receive multidisciplinary treatment with curative intent.*

$\Downarrow$ refined_as

(2) {*Recommendation*}*:* {*Patients with*} [disease] {*should receive*} [treatment] {*with*} [med_goal].

$\Downarrow$ refined_as

(3) {*Recommendation*}: [Target_group] [recommendation_op] {*receive*} [treatment] {*with*} [med_goal].

$\Downarrow$ refined_as

(4) {*Recommendation*}: [med_context] [recommendation_op] [complex_treatment].

Figure 4 Abstraction steps for extracting a pattern template.

(which contains relations such the ones shown in Fig. 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. In this case, *med_recommendation* becomes a component that encodes the different regularities representing a medical advice, which are all recognized using the *recommendation_op* operator (a class of lexical markers, such as ''should'').

The goal of finding linguistic templates in the text requires finding of 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 generated and instantiated automatically using this method, producing meaningful linguistic pattern templates that are medically relevant cannot be fully automated, but requires manual selection. This selection of basic medical knowledge templates is depicted as step 4 of the procedure in Fig. 2 and corresponds to activity 2 in Fig. 1.

3.3. Detection of pattern instances in the guideline text

In this section we discuss step 4 of the algorithm in Fig. 2. For identifying instantiations of pattern templates in the guideline text we use two custom built ontologies: one of the medical domain, and one reflecting the control aspect allowed by the target guideline modelling language. For the sake of simplicity, we will refer to these two ontologies as being one ontology. Fig. 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).

Please cite this article as: Radu Serban et al., Extraction and use of linguistic patterns for modelling medical guidelines, Artificial Intelligence in Medicine (2006), doi:10.1016/j.artmed.2006.07.012

To decide which concepts are present in our ontology, we used the Text2Onto tool [29] to extract the most frequently used medical terms from a corpus of guideline text, then categorized these terms according to semantic types present in UMLS thesaurus. Other ontology extraction methods from a corpus of text have recently been used to mine lists of frequent terms in an unannotated corpus [30]. From the text fragments in which the most frequent terms occurred, we selected constructs corresponding to medical relations, such as: *Therapy A* **helps against** *disease B*. *Treatment A* **consists of** *therapies B,C,D*. *Drug A* **helps against** *disease B*, etc. These relations form the core of our guideline ontology, which is an extended medical domain model. We then established mappings between these constructs (or medical categories of their components) and semantic types in UMLS, and imported the UMLS relations associated with those semantic types into our custom ontology. Based on such mappings, UMLS relations such as *ClinicalDrug* **affects** *Body_Location_or_Region* are transformed into relations in our guideline ontology, such as *Medication* **affects** *BodyPart*.

An application we built then generates templates (i.e., knowledge placeholders) as sequences of slots associated with medical concepts from a specific category, connected by medical relations allowed by the guideline modelling domain, and instantiates them, for instance:

instance([*radiotherapy,produces,skin_reactions*]) **instance_of** *template* ([*med_action,effect_op,med_effect*]) **covers** *ontology_fragment* (*MedAction produces MedEffect*).

We defined 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. By coupling the knowledge templates with these control relations, we are able to select control templates and instantiate them in the guideline text.

The guideline text is split into sentences and 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*. The process of pattern detection is an iteration of several semantic tagging and pattern recognition steps, in which the chunk is initialized at the word level, then after semantic annotation using background knowledge the chunks corresponding to several words making up sensible medical terms or medical sentences are merged, depending on the level of granularity at which patterns are recognized.

A pattern (template) can be viewed as 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 that can be connected with relations in a guideline representation language. We define patterns at different levels of granularity: (1) patterns at word-level are in fact semantically tagged medical terms in the guideline text (multiple words are grouped according to a custom heuristic based on a dictionary lookup); (2) pattern at sentence level define concepts from different semantic categories which correspond to well-defined formal constructs.

After splitting the guideline into word-level chunks, the list of annotations of each word contains only the relative position of the term in the guideline text. An iterative annotation takes place, first within sentence, then within larger fragments. At each processing of a set of chunks, 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 within a sentence, the chunks corresponding to the component words are merged into one chunk, together with their annotations. For finding overlapping patterns, the analysis focuses on sentence-level chunks, which are sequences of word-level chunks found within a sentence border. The result of this step is annotating each sentence with all template instances found within that sentence.

3.4. Selection of core patterns

In order to determine components that are useful in modelling the guideline, we have to establish the set of "atomic" templates which produce minimal models, by looking at the relations between templates. After detecting the instances of medically-relevant linguistic templates in the guideline text, we choose as basic pattern templates those which have the highest support and are more abstract than other templates.

Definition 1. A linguistic template *LT* is an alternating list of domain-specific $\mathrm{d}t_i$ and control relation expressions ct_i, possibly prefixed with lexical literals lt_k: $LT = \langle lt_0, \mathrm{d}t_0, lt_1, ct_1, lt_2, \mathrm{d}t_1 \rangle$, where lt_k can be the empty string, or can match control expressions; $\mathrm{d}t_i$ is restricted by the vocabulary allowed by the domain ontology; and ct_i is restricted by the vocabulary allowed by the control ontology.

Please cite this article as: Radu Serban et al., Extraction and use of linguistic patterns for modelling medical guidelines, Artificial Intelligence in Medicine (2006), doi:10.1016/j.artmed.2006.07.012

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 $BK = DO \cup CO$ 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 $S_1 \in S_{Cat}$. Linguistic templates with a high level of abstraction represent maximal schemas. For selecting the core templates, we define relations between linguistic templates $LT_1 = [C_{11}, C_{12}, \ldots, C_{1n}]$ and $LT_2 = [C_{21}, C_{22}, \ldots, C_{2n}]$, using the hierarchical relations in the domain + control ontology:

is-more-specific $(LT_1, LT_2)]$ *iff* for all $i = \overline{1, n}$: is_a(C_{1i}, C_{2i});
contains $(LT_1,\ LT_2)]$ *iff* $\{C_{21}, C_{22}, \ldots C_{2n}\} \subset \{C_{11}, C_{12}, \ldots, C_{1n}\}$.

More generic connections between templates can be established:

is-similar (LT_1, LT_2) if *contains* (LT_1, LT_3) and *is-more-specific* (LT_3, LT_1).

These mappings can be aided by using categorization of the lexical markers present in the templates. The process of pattern instance detection produces a list of pattern templates and a lexicon of link words that connect medical terms in the pattern instances detected. For the guideline analyzed [21], 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

These lexical markers help us recognize the linguistic patterns in the text. If two sentences use two different *recommendation_op* markers (should, advisable), they are more likely to be recognized as being composed of recommendation templates which have a standard modelling schema in the formal representation of the guideline. In some cases, these markers correspond to semantic relations in the ontology of the guideline representation domain: ordering of actions, quantification of action effects, etc. By mapping the linguistic templates to control structures allowed by the guideline representation language, one is able to define a modelling schema for the template.

Definition 2. A linguistic (knowledge transformation) pattern LP is a mapping $\langle LT, MapR, CT \rangle$ between a linguistic template LT and its corresponding control translation CT in a guideline modelling language, using a set $MapR$ of mapping rules between elements of the template and elements of the formal translation.

A list of the core templates with a high support among the instances identified in our reference guideline is summarized in Table 1, along with the modelling schema of each template and with frequencies of occurrence in the three guideline chapters used in the evaluation in Section 4. These pairs, plus the semantic mark-up rules MapR, make up the triples seen in the definition above.

Table 1 Coverage of core pattern templates in the chapters analyzed

Core template name, categories instance example	Translation, freq. (ch. 2—4)
Association action-goal: [*med_action, assoc_rel_op, disorder*] [*surgery, to_reduce, tumour_load*]	*Action—goal*, 10 *occurrences*
Action decomposition: [*med_action, decomp_op,med_action, med_action*] [*current_treatment, consists_of, surgery, radiotherapy*]	*Decomposition*, 3 *occurrences*
Association condition-action: [*med_context, med_action*] [*multidisciplinary_treatment, chemotherapy*]	*If—then*, 12 *occurrences*
Action sequencing: [*med_action, act_op, med_action*] [*radiotherapy, following, neoadjuvant_chemotherapy*]	*Sequencing*, 29 *occurrences*
Associations action-effect: [*disorder, temp_rel_op, med_action*] [*tumour_recurrence, following, radiotherapy*]	*Action—effect*, 2 *occurrences*
Preference for actions: [*treatment, assoc_rel_op, med_action*] [*treatment_of_choice, is, neoadjuvant_chemotherapy*]	*Preferences*, 19 *occurrences*

Please cite this article as: Radu Serban et al., Extraction and use of linguistic patterns for modelling medical guidelines, Artificial Intelligence in Medicine (2006), doi:10.1016/j.artmed.2006.07.012

3.5. Mapping the core templates into formal constructs

The parameters of pattern instances detection are: (1) the combined medical + control domain ontology; and (2) a set of target pattern templates sought in the text. After applying the algorithm described above to the reference guideline [21], and reviewing the instances found, the most frequent operational patterns were:

$p_{1.1}$ (*A*: *med _action*, {*following*}: *seq _act _op*, *B*: *med _action*);
$p_{1.2}$ (*A*: *med _action*, {*after*}: *seq _act _op*, *B*: *med _action*);
$p_{1.3}$ (*A*: *med _action*, {*consists _of*}: *decomp _op*, *B*:*med _action*, *C*: *med _action*).

The first two items are subclasses of a more abstract pattern—sequence of two medical actions, denoted: *p1* (*med _action, seq _act _op, med _action*). In Fig. 5 we have depicted a few templates *p*1 corresponding to pattern instances *i*1, *i*2. Pattern *p*1 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 of such pattern instances were identified.

By grouping together the template instances which are similar or share common words, the most frequent linguistic constructs can be refined and then used as in building blocks for 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. In the case above, an action-sequencing transformation *DO* (*medical _action* [1])*AFTER DO* (*medical _action* [2]) for p1 produces:

{*DO* (*excision*); *DO* ({*biopsy, axillary _surgery*})};
{*DO* (*mastectomy*); *DO* (*breast _reconstruction*)}.

We have summarized in Table 1 the most frequently used transformations of linguistic templates encountered in the reference guideline fragment analyzed into generic elements of the ASBRU guideline representation language.

4. Evaluating the use of patterns in guideline formalization

Guideline formalization is a transformation that takes as input a guideline text GL and a set of formalization rules RF, and produces an executable representation E of the procedural part of the guideline. Our linguistic pattern-driven approach to formalization consists in deriving a set of constraints RF by reverse-engineering, using a domain-specific lexicon, of the mappings between text fragments and medical procedural knowledge, and using the representation of that knowledge in the guideline representation language to obtain E. 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. To evaluate how close two executable models are,

```
i1(axillary_surgery, following, excision)
i2(breast_reconstruction, following, mastectomy)
```

⇓ instance_of

p1(med_action[1], seq_act_op, med_action[2])

```
i3(treatment, of, locoregionally_advanced_breast_cancer,
   consists_of, neoadjuvant_chemotherapy, followed_by,
   surgery, and, locoregional_radiotherapy)
```

⇓ instance_of

p2(med_action[1], rel_op, med_disease)∘

p3(med_action[1], decomp_op, med_action[2], seq_act_op, med_action[3])

where $med\_action[3] :=$ **p4(med_action[4], act_op, med_action[5])**

Figure 5 Pattern templates extracted from instances.

Please cite this article as: Radu Serban et al., Extraction and use of linguistic patterns for modelling medical guidelines, Artificial Intelligence in Medicine (2006), doi:10.1016/j.artmed.2006.07.012

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.

4.1. Evaluation results

We have compared the results of modelling chapters 2, 3 and 4 of the **CBO guideline for treatment of breast cancer**[21] in the intermediate representation MHB [13,22], 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. MHB [13] was selected as intermediate guideline representation language, because it supports the control constructs allowed in ASBRU, is general enough to support other target guideline representation languages, and can be used to validate a semi-formal representation of a guideline by medical experts and by knowledge engineers.

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.

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. We focused on improving the recall of relevant sentences containing procedural knowledge. 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 sentences 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. For instance, in chapter 2, from the 130 sentences that have been selected automatically by the pattern detection application (out of a much larger number of sentences), only 30 were relevant for the model produced manually by the knowledge engineer. This indicates a recall of 30 correct markings out of 41 marked up by the knowledge engineer, i.e. 73% recall with respect to the golden standard input. The linguistic templates instantiated in chapter 2 were translated into candidate semi-formal constructs in MHB, but only 8 of them were included ad-literam in the golden standard MHB model produced manually, i.e. a precision of 19.5% with respect to the MHB model. A measure of the effectiveness of the automatic translation from text to MHB is given in terms of the number of relevant MHB constructs generated, in relation to the number of relevant text constructs marked up automatically (the rate of 23% in the last column). This rather low precision and effectiveness is due to the fact that not all sentences contributed in the same manner to the model and some additional semantic interpretation steps were performed manually, which could not be done by the application. Furthermore, the human modeller mapped

Table 2 Evaluating the effectiveness of linguistic pattern detection in guideline modelling: selection of relevant sentences using automated mark-up and linguistic pattern detection vs. manual annotation

Sentences	(Autom.) processed	(Manual.) modelled	Modelled and processed (recall)	Modelled and processed and with relevant 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%)

Please cite this article as: Radu Serban et al., Extraction and use of linguistic patterns for modelling medical guidelines, Artificial Intelligence in Medicine (2006), doi:10.1016/j.artmed.2006.07.012

more than one sentence to the same MHB construct, which was not the case with the MHB model generated automatically, which generated one candidate (typically, too finely grained to be considered equivalent to a MHB fragment modelled by the knowledge engineer) for each pattern 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 or to terms absent from the ontology, which could not be extracted. Better coverage heavily depends on having a complete classification of medical terms, particularly actions. Using a richer domain ontology and especially a more elaborated control ontology for detecting patterns used in generating the executable skeleton of the guideline model would prove helpful in supporting formalization.

5. Related work

Guideline patterns reflect modelling decisions when medical guidelines are transformed into an executable form. The existing guideline frameworks, such as EON [23], DEGEL [24] or GUIDE [25], employ medical vocabularies, vocabulary servers, or are concerned with the role of semantic mark-up in representing medical guidelines, but either do not address at all, or make no clear reference to the semantic mark-up as part of guideline formalization. For extracting structure and semantics from annotated and unannotated text, to support querying and text summarization, we benefit from existing Natural Language Processing techniques applied for text and data mining (see, for instance, MedLEE [26], MedSyndicate [9,27] and other similar work). Information Extraction relies on syntactic and semantic tagging of plain text [28,30—33], in order to extract syntactic constructs, vocabularies, or even ontologies. The tagging is performed using background knowledge in the form of a dictionary, thesaurus, positive examples of mappings, or conceptual graphs [34,35]. Statistical and probabilistic models [28,36] were used to increase the performance when ambiguous textual constructions are present. More recently, Rindflesch and Fiszman [1] proposed a methodology of combining domain knowledge with linguistic structure for facilitating interpretation of context citations in medical texts. Our work has similarities with concept and relation extraction [37—39] and with semantic mark-up and interpretation of medical texts [40,24], but focuses on the use of an ontology to generate and validate knowledge transformation patterns for medical texts obeying rather strict formatting rules. The limitations of NLP techniques [41] in tasks such as entity recognition, term disambiguation, relation extraction through syntactic analysis, are also present in our approach and need to be addressed for a better performance of our method. Despite of its limitations, our proposed approach to guideline formalization, using semantic mark-up along a domain + control ontology and linguistic patterns translation to guideline model fragments, provides a potential effort reduction in the guideline formalization process.

6. Conclusions

Searching of 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 guidelines analyzed. Linguistic patterns are basic building blocks from which semantically-richer fragments can be built, facilitating modularization, validation and reuse of the background knowledge covered by guidelines.

We introduce a method to extract control knowledge from the text of medical guidelines, by instantiating and translating automatically one or more predefined linguistic patterns. This step can be performed in the initial phase of guideline formalization, as it guides the manual modelling of guidelines by a knowledge expert and leads to a reduction of the effort spent in modelling guidelines. We provide an initial evaluation of the usefulness of our method, by measuring the precision of detecting the procedural knowledge used in guideline formalization, and the coverage of the gold standard model.

The search for linguistic patterns useful in guideline formalization is guided by the mappings between the medical terms occurring in guidelines and the concepts in a medical ontology. These mappings 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.

The use of patterns produces a lexicon and a skeleton of the formal model covered by the procedural part of the guideline, automatically. Therefore, the best use of a pattern extraction tool such as the one described above should be coupled with a semantic and structuring guideline mark-up tool (see [24,3,12]), which has already delineated the procedural part of the guideline. The method proposed for extracting candidate patterns can be extended to non-procedural knowledge, therefore authoring and formalization of medical guidelines can benefit from the use of this ontology-driven approach to obtaining linguistic patterns. We propose the use of our method as a pre-processing step that assists, and does not replace, the role of the knowledge engineer in the guideline formalization process. By proposing an automatic translation of the medical terms conforming to linguistic patterns, into a more formal representation in one of the guideline representation languages, it reduces the cognitive load for the knowledge engineer, allowing him/her to concentrate on less regular knowledge which is more difficult to interpret.

Acknowledgment

This work has been supported by the European Commission's IST program, under contract number IST-FP6-508794 Protocure-II: http://www.protocure.org. Accessed: 1 June 2006.

References

[1] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003;36(6):462—77.
[2] Protocure 2 Project. Integrating formal methods in the development process of medical guidelines and protocols. URL: http://www.protocure.org. Accessed: 1 June 2006.
[3] Svatek V, Ruzicka M. Mark-up based analysis of narrative guidelines with the Stepper tool. In: Kaiser K, Miksch S, Tu SW, editors. Proceedings of Symposium on Computerized Guidelines and Protocols (CGP-04). Prague, Czech Republic: IOS Press; 2004. p. 132—6.
[4] Moser M, Miksch S. Improving clinical guideline implementation through prototypical design patterns. In: Keravnou E, Miksch S, Hunter J, editors. Proceedings of 10th Conference on Artificial Intelligence in Medicine (AIME 2005). Aberdeen, UK: Springer-Verlag GmbH; 2005 July . p. 126—30 [LNCS 3581].
[5] Uren V, Buckingham S, Mancini C, Li G. Modelling naturalistic argumentation in research literatures. In: Proceedings of 4th Workshop on Computational Models of Natural Argument, Valencia, Spain, 22—27 August; 2004.
[6] Ruch P, Rassinoux AM, Baud RH, Lovis C. A light knowledge model for linguistic applications. In: Proceedings of American Medical Informatics Association Symposium; 2001. p. 37—41.
[7] Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: Proceedings of American Medical Informatics Association Symposium. National Library of Medicine; 2001. p. 17—21.
[8] Rindflesch TC, Aronson AR. Semantic processing for enhanced access to biomedical knowledge. In: Kashyap V, Shklar L, editors. Real World Semantic Web Applications. IOS Press; 2002. p. 157—72.
[9] Hahn U, Schulz S, Romacker M. How knowledge drives understanding—matching medical ontologies with the needs of medical language processing. Int J Artif Intell Med 1999;15(1):25—51.
[10] Balser M, Coltell O, van Croonenborg J, Duelli C, van Harmelen F, Jovell A, et al. Protocure: supporting the development of medical protocols through formal methods. In: Kaiser K, Miksch S, Tu SW, editors. Proceedings of Symposium on Computerized Guidelines and Protocols (CGP-04). Prague, Czech Republic: IOS Press; 2004 April.
[11] Bury J, Ciccarese P, Fox J, Greenes RA, Hall R, Johnson PD, et al. Comparing computer-interpretable guideline models: a case-study approach. J Am Med Inform Assoc 2003;10(1): 52—68.
[12] Seyfang A, Kosara R, Miksch S, Votruba P. Tools for acquiring clinical guidelines in ASBRU. In: Proceedings of the 6th World Conference on Integrated Design and Process Technology (IDPT'02); 2002.
[13] Seyfang A, Miksch S, Marcos M, Wittenberg J, Rosenbrand K, Polo-Conde C. Bridging the gap between informal and formal guideline representations. In: Proceedings of the 17th European Conference on Artificial Intelligence (ECAI 2006); 2006.
[14] Beckworth R, Fellbaum C, Gross D, Miller G. WordNet: A lexical database organized on psycholinguistic principles. In: Zernik U, editor. Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon. Lawrence Erlbaum Associates; 1991. p. 211—26.
[15] Spackman KA, Campbell KE, Cote RA. SNOMED RT: A reference terminology for health care. J Am Med Inform Assoc; 1997. p. 640—4 [Fall Symposium Supplement].
[16] Rector A, Gangemi A, Galeazzi E, Glowinski A, Rossi-Mori A. The GALEN CORE model schemata for anatomy: towards a re-usable application-independent model of medical concepts. In: Proceedings of the Twelfth International Congress of the European Federation for Medical Informatics in Europe (MIE 94); 1994. p. 229—33.
[17] Mesh (Medical Subject Headings (MESH)). URL: http://www.nlm.nih.gov/mesh/meshhome.html. Accessed: 1 June 2006.
[18] Unified Medical Language System (UMLS). URL: http://www.nlm.nih.gov/research/umls/. Accessed: 1 June 2006.
[19] National Cancer Institute (NCI) ontology. URL: http://www.mindswap.org/2003/CancerOntology/. Accessed: 1 June 2006.
[20] Johnson S. A semantic lexicon for medical language processing. J Am Med Inform Assoc 1999;6(3):205—18.
[21] Dutch Institute for Healthcare Quality (CBO), Guideline for the Treatment of Breast Carcinoma; 2002. PMID: 12474555.
[22] Seyfang A, Miksch S, Polo-Conde C, Wittenberg J, Marcos M, Rosenbrand K. MHB—a many-headed bridge between informal and formal guideline representations. In: Miksch S, Hunter J, Keravnou E, editors. Proceedings of 10th Conference on Artificial Intelligence in Medicine (AIME 2005). Springer; 2005. p. 146—50.
[23] Tu SW, Musen MA. A Flexible Approach to Guideline Modeling. In: Nancy Lorenzi P, editor. 1999 AMIA

Please cite this article as: Radu Serban et al., Extraction and use of linguistic patterns for modelling medical guidelines, Artificial Intelligence in Medicine (2006), doi:10.1016/j.artmed.2006.07.012

Annual Fall Symposium. Washinton D.C.: Hanley & Belfus Inc.; 1999 . p. 475—97.

[24] Shahar Y, Young O, Shalom E, Mayaffit A, Moskovitch R, Hessing A, et al. DeGeL: A Hybrid, Multiple-Ontology Framework for Specification and Retrieval of Clinical Guidelines. In: Dojat M, Keravnou ET, Barahona P, editors. Proceedings of 9th European Conference on AI in Medicine (AIME 03). Heidelberg: Springer-Verlag; 2003. p. 122—31.

[25] Ciccarese P, Caffi E, Boiocchi L, et al. The NewGuide Project: guidelines, information sharing and learning from exceptions. In: Dojat M, Keravnou ET, Barahona P, editors. Proceedings of 9th European Conference on AI in Medicine (AIME 03), Heidelberg: Springer-Verlag; 2003.

[26] Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. In: CC Chute, editor. Proceeding Conference on Natural Language and Medical Concept Representation; 1997. p. 41—52, http://citeseer.ist.psu.edu/friedman97evaluating.html.

[27] Hahn U, Romacker M, Schulz S. medSyndikate: a natural language system for the extraction of medical information from findings reports. Int J Med Inform 2000;67(1/3):63—74.

[28] Paynter GW, Witten IH, Gutwin C, Frank E, Nevill-Manning C. Domain-specific keyphrase extraction. In: Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers; 1999p. 668—73.

[29] Cimiano P, Volker J. Text2Onto: A framework for ontology learning and data-driven change discovery. In: Metais E, Montoyo A, Munoz R, editors. Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005. Springer; 2005. p. 227—38 [LNCS].

[30] Hernandez CP, Ortiz AM. From text to ontology: extraction and representation of conceptual information. In: Proceedings of the 4th Conference Terminology and Artificial Intelligence, Nancy, France; 3—4 May 2001.

[31] Riloff E. Automatically generating extraction patterns from untagged text. In: Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96); 1996. p. 1044—9.

[32] Riloff E, Shoen J. Automatically acquiring conceptual patterns without an annotated corpus. In: Yarovsky D, Church K, editors, Proceedings of the Third Workshop on Very Large Corpora. Somerset, New Jersey: Association for Computational Linguistics; 1995. p. 148—61.

[33] Huffman SB. Learning information extraction patterns from examples. In: Wermter S, Riloff E, Scheler G, editors. Connectionist, statistical, and symbolic approaches to learning for natural language processing. Berlin: Springer; 1995. p. 246—60.

[34] Witten I. Adaptive text mining: inferring structure from sequences. J Discrete Algorithms 2004;2(2):137—59.

[35] Zhong J, Zhu H, Li J, Yu Y. Conceptual graph matching for semantic search. In: Proceedings of the 10th International Conference on Conceptual Structures: Integration and Interfaces. Springer-Verlag; 2002. p. 92—106.

[36] Culotta A, McCallum A, Betz J. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In: Proceedings of the Human Language Technology Conference, HLT-NAACL 2006. New York City: Association for Computational Linguistics; 2006 June .

[37] Ceusters W, Buekens F, de Moor G, Waagmeester A. The distinction between linguistic and conceptual semantics in medical terminology and its implication for NLP-based knowledge acquisition. Methods Inform Med 1998;37(4—5):327—33.

[38] do Amaral MB, Roberts A, Rector AL. NLP techniques associated with the OpenGALEN ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logical constructs. In: Overhage JM, editor. Proceedings of the 2000 American Medical Informatics Association Annual Symposium (AMIA 2000). Los Angeles: American Medical Informatics Association; 2000. p. 76—80.

[39] Leroy G, Chen H. GeneScene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. J Am Soc Inform Sci Technol 2005;56(5):457—68.

[40] Deneke K, Kohlhof I, Bernauer J. Use of multiaxial indexing for information extraction from medical texts. In: Proceedings of the Workshop on Foundations of Clinical Terminology and Classifications (FCTC 06), Timisoara, Romania. ROMEDINF; April 2006.

[41] Ciravegna F. Adaptive Information Extraction from Text by Rule Induction and Generalization. In: Nebel B, editor. Proceedings of the Seventeenth International Conference on Artificial Intelligence (IJCAI-01). San Francisco, CA: Morgan Kaufmann Publishers, Inc.; 2001. p. 1251—6.

Please cite this article as: Radu Serban et al., Extraction and use of linguistic patterns for modelling medical guidelines, Artificial Intelligence in Medicine (2006), doi:10.1016/j.artmed.2006.07.012