Supporting the Abstraction of Clinical Practice Guidelines Using Information Extraction

Katharina Kaiser[1,2] and Silvia Miksch[2,3]

[1] Center of Medical Statistics, Informatics & Intelligent Systems
Medical University of Vienna
Spitalgasse 23, 1090 Vienna, Austria
[2] Institute of Software Technology & Interactive Systems
Vienna University of Technology
Favoritenstraße 9-11/188, 1040 Vienna, Austria
[3] Department of Information & Knowledge Engineering
Danube University Krems
Dr.-Karl-Dorrek-Straße 30, 3500 Krems, Austria

Abstract. Modelling clinical practice guidelines in a computer-interpretable format is a challenging and complex task. The modelling process involves both medical experts and computer scientists, who have to interact and communicate together. In order to support both modeller groups we propose to provide them with helpful information automatically generated using NLP methods. We identify this information using rules based on both syntactic and semantic information. The majority of the defined information extraction rules are based on semantic relationships derived from the UMLS Semantic Network. Findings in the evaluation indicate that using rules based on semantic and syntactic information provide valuable and helpful results.

1 Introduction

Clinical practice guidelines (CPGs) are important instruments to provide state-of-the-art clinical practice at point of care for the medical and clinical personnel [1]. They typically address a specific health condition and provide recommendations to the physician about issues such as who to investigate for the problem, how to investigate it, how to diagnose it, and how to treat it. It has been shown that integrating CPGs into clinical information systems can improve the quality of care [2]. For integrating CPGs into such systems, the CPGs (i.e., documents in free, narrative text) have to be translated into a computer-interpretable format. Various of such guideline representation formats have been developed (see [3] for an overview and comparison), but the translation process is still a large bottle-neck. It's a complex and challenging task requiring both medical and computer science expertise. Several methods have been developed that describe systematic approaches for guideline modelling. Some of them use intermediate formats to break the complexity of the modelling task into manageable tasks.

In various projects we use the *Many-Headed Bridge between guideline formats* (MHB) [4] as intermediate abstraction format. But still the modelling, involving both

medical experts and computer scientists, is challenging and requires high training effort. Thus, we want to support the abstraction process by automatically identifying information dimensions regarding the MHB format using NLP. Our overall goal is to accomplish the abstraction process almost automatically and generate the MHB model forthwith. As a first step, we want to provide identified information dimensions to the modellers to support the further assignment to more detailed information slots.

In this paper we describe methods and an application to automatically identify information dimensions necessary for modellers of computer-interpretable guidelines using NLP.

In the next section we give a short overview on guideline modelling, knowledge-based methods for support, and describe MHB's information dimensions. In Section 3 we explain methods developed and resources used. Section 4 describes the evaluation of our system and a discussion of its results. The final section contains our conclusions.

2 Background

Modelling CPGs in a computer-interpretable and -executable format is a great challenge. Different actors are involved in this task (i.e., medical experts and computer scientists) and must find a way to communicate together.

For non-medical scientists the language in such documents is difficult to understand. On the one hand they are unfamiliar with the medical concepts, on the other hand the medical language implies a different meaning than generally assumed. For instance, the sentence *'Neonatal respiratory depression could be relieved by naloxone administration.'* means that under the condition "neonatal respiratory depression" one should perform the activity "naloxone administration" with an effect "relief". Understanding the medical language in CPGs and differentiating between explanatory information and activities to be performed are therefore the major challenge for computer scientist modellers.

As modelling is similar to programming medical scientists have difficulties with the formal concepts. Furthermore, the medical language is often vaguely formulated. To bring medical scientists to formulate concrete statements is a major challenge. In principle, MHB supports the formalization by keeping expressions used in the CPG. This makes the modelling easier for medical scientists, but still they have difficulties in detecting parameters needed and ordering of tasks.

2.1 Knowledge-based Methods for Guideline Modelling

Several attempts have been made to support the modelling, maintenance, and shareability of computer-interpretable guidelines. For instance, Serban et al. defined linguistic patterns found in CPGs that can be used to support both the authoring and the modelling process of guidelines [5]. Furthermore, they defined an ontology out of these patterns and linked them with existing thesauri in order to use compilations of thesauri knowledge as building blocks for modeling and maintaining the content of a medical guideline [6]. In [7] we proposed a method to identify actions in ontolaryngology CPGs using a subset of the *Medical Subject Headings* (MeSH) as a thesaurus.

2.2 The *Many-Headed Bridge* between Guideline Formats

The MHB format [4] is a XML language aimed at representing an abstract representation of a CPG while translating it into the target representation language. It was designed to close the large gap between the natural language text and the formal representation of a clinical guideline, thus facilitating the formalization process. MHB provides chunks which correspond to a certain bit of information in the natural language text (e.g., a sentence, part of a sentence, or more than one sentence). These chunks are internally structured into aspects which are grouped by eight dimensions: control flow, data flow, temporal aspects, background information, evidence, resources, patient related aspects, and document structure. The most complex are the first four dimensions, which we will describe more detailed. Each dimension is then further divided into more specific information slots.

1. **Control Flow** specifies the execution order of tasks, their decomposition, and the gathering of information. It can be used to specify decisions, ordering, decomposition, and synchronization of tasks as well as actions.
2. **Data Flow** describes the data processing involved in the diagnosis and treatment.
3. **Temporal Aspects** of qualitative or quantitative nature can be specified.
4. **Background Information** can motivate the reader to follow the guideline or information to complement the statements in the recommendation part. That can be intentions, effects, relations, relations, educational information, explanations, and indicators.

Our methods are tailored to these four dimensions, which are the most challenging to be identified by modellers and also the most important ones for guideline modelling, because they contain the minimal requirements for executing guideline processes.

3 Methods

In the ReMINE project[4] we modelled a local adaptation of a guideline for natural birth for an Italian hospital. It was derived from the guideline “Induction in labour” [8]. Resulting from difficulties during the modelling process we started using the document for searching for patterns that could be used to identify several MHB dimensions. In order to generate patterns that work for our purpose, we had to use both syntactical and semantic information from the text. We defined hand-written rules to identify conditions and actions for the control dimension, data items (i.e., parameters) needed for processing conditions for the data flow dimension, temporal aspects for the time dimension, and background information.

We use the MetaMap Transfer Program (MMTx) [9] to identify the medical concepts and their semantics in the text, the Stanford Parser [10] to generate a parse tree, and its utility Tregex [12] for matching patterns in trees.

[4] http://www.remine-project.eu; last accessed: Jan. 12, 2010

3.1 Identifying Instances of Control Dimensions

The most challenging part of modelling is the detection of actions or activities. How can we differentiate between an action or activity and background information? The UMLS Semantic Network [11] offers amongst various semantic types also semantic relationships between these types and therefore acts as an upper-level ontology. We use these semantic relationships to identify actions.

We analysed actions in our CGP and found both complete and incomplete relationships. The latter result from the fact that actions and activities in CPGs are related to tasks the health care personnel has to perform (in our specific case: gynaecologists and obstetricians, who are assigned the semantic type *Professional or Occupational Group*). As most of the actions address these users directly, they are only implicitly referred to in the text. Furthermore, passive format is frequently used, which also allows omission of the agent of the action.

In the Semantic Network there are 61 left-hand side semantic relations defined for *Professional or Occupational Group*. We use 18 of these relations that are related to activities performed by health care personnel, such as *'***treats** *Patient or Disabled Group'*, *'***performs** *Therapeutic or Preventive Procedure*', and so on, and combined them to five relations of types **diagnoses**, **interacts_with**, **performs**, **treats**, and **uses**. Furthermore, when we analysed the semantic relationships in our text, we found also new relationships that we added (e.g., **acquires** *Finding*).

The more challenging actions are those that formulate the activity implicitly. We identified 27 complete relations referring to actions, such as *'Therapeutic or Preventive Procedure* **treats** *Disease or Syndrome'*, *'Patient or Disabled Group* **performs** *Therapeutic or Preventive Procedure'*, or *'Therapeutic or Preventive Procedure* **affects** *Patient or Disabled Group'*, which use relation types **treats**, **uses**, **performs**, and **affects**. For instance, the sentence 'Women may eat a light diet in established labor.' can be described by the relation *'Population Group* **performs** *Therapeutic or Preventive Procedure'*.

In order to define patterns that indicate such semantic relationships we were confronted with the challenge of identifying the type of relationship. In most cases a specific verb suggests the relation type. Therefore, we tried to identify verbs for each type of relationship. We analysed our training document and assigned each relation type the verbs appearing. Afterwards we also collected synonyms of these verbs in online dictionaries. Thereby, we were able to generate syntactic rules based on semantic information. Figure 1 shows the implementation of the semantic relationships in *tregex* patterns [12] and gives examples how they work. We are now able to omit sentences not describing actions.

Actions and activities can be controlled by conditions. In many cases conditions start with 'if', 'in case of', 'when', and so on. For this kind of conditions we easily generated syntactic patterns. Other conditions could be identified based on their semantic types. For instance, the sentence 'Women with pain but no cervical changes should be reexamined ...' contains the condition 'with pain but no cervical changes'. 'Pain' is assigned the semantic type *Sign or Symptom* and 'cervical changes' is assigned *Finding*. Thus, we defined condition patterns based on semantic information, for instance, based on semantic relation *'Finding* **occurs_in** *Population Group'*.

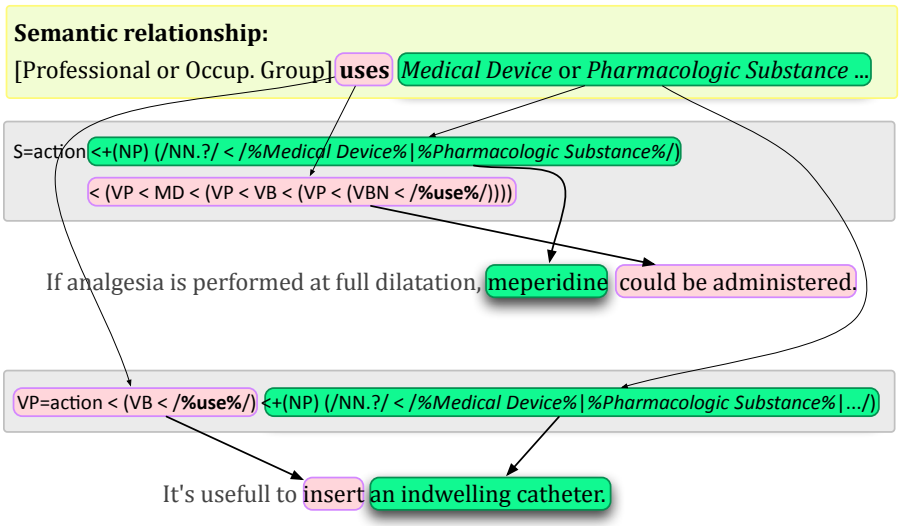

Fig. 1. *tregex* patterns implementing the relationship '[Professional or Occupational Group] **uses** *Medical Device* or *Pharmacologic Substance* or ...'.

3.2 Identifying Instances of Background Dimensions

The background dimension comprises different kinds of information, such as explanatory information, intentions, effects, and so on. We formulated patterns that are solely based on syntactic information. Thereby, expressions such as 'even if ...', 'in order to ...', 'because ...', or 'although ...' are used to identify different aspects of background information.

3.3 Identifying Instances of Time and Data Dimensions

Identifying temporal aspects in a CPG is a very important issue, as we can thereby suggest for connections and dependencies among actions. We generated patterns based on syntactic information and regular expressions.

Identifying data dimensions within a text has emerged to be difficult for both computer scientists and medical scientists. The data dimension refers to data items that are read from an electronic health record or asked from the personnel (i.e., 'usage'), data items that receive a certain information (e.g., newly generated information that is written into the electronic health record) (i.e., 'input'), and data items that are computed by some rules, which are defined in the document (i.e., 'abstraction'). The first ones are occurring within condition clauses, the second ones within actions, and the latter ones are appearing independently in the text. We generated patterns for input and usage of data using semantic information. Thereby, 'input' patterns are associated with certain 'action' patterns. For instance, [*Professional or Occupation Group*] **acquires** *Finding*

or *Organism Function* or *Body Substance* ... also indicates for data input, which is contained in the right-hand side of the relationship.

3.4 Providing the Results

In order to provide the obtained information to the users in a simple and clearly represented way, we decided to generate a simple HTML output highlighting the different information dimensions with different colors. Due to the small amount on information dimensions the representation should be comprehensible; minimal learning effort is required. Modelling is accomplished using the DELT/A tool [13]. Thereby, the HTML document is loaded. By marking up the various dimensions the corresponding MHB model is built. Using the highlighted HTML output has the advantage that it is also displayed in the DELT/A tool and the highlighting can be used immediately for supporting the users.

4 Evaluation

We implemented a prototype system that processes the defined patterns and generates the HTML output described above. We then applied the system to another document to evaluate the performance of our system.

We used the guideline “Management of labor” [14] for evaluation. Although the CPG covers the same application area, it is structured in a different way, uses different phrasing and wording. Furthermore, it contains lots of background information and literature references and it is almost twice as large (60 pages) as the document we used to define our rules.

Due to the lack of a gold standard, we manually checked the output of our program with the help of an MHB expert. Table 1 shows the evaluation results. The overall recall and precision values are 69.3% and 79.3%, respectively. As we had expected a better result, we analysed how this outcome arose.

Our analysis showed that most missing assignments of actions base on incorrect parsing information and not detecting the appropriate semantic relation either due to a verb not incorporated in the knowledge base or due to not detecting the semantic type of a phrase. Incorrectly assigned actions result from wrong assignments of semantic types. The largest influence on the overall result has the erroneous output of the parser.

A minor source of errors are the completeness of semantic relationships and verbs pointing to a relation. We think that expanding our patterns by analysing more documents will improve the output.

We also provided the output of the system to students modelling the document in MHB. First feedback from medical informatics students shows that by providing information on dimensions the modelling is easier and faster accomplished. The resulting models are significantly more elaborated than those without this additional information. We are now performing experiments, for instance, to get insights which of the linguistic tasks are more critical to effectively support CPG modelling.

Table 1. Evaluation results.

	CONTROL		TIME	DATA	BACKGR.	Overall
	Condition	Action				
COR	40	88	29	55	117	**329**
INC	20	27	5	11	17	**80**
PAR	5	2	5	2	2	**16**
POS	79	150	39	74	144	**486**
ACT	65	117	39	68	136	**425**
REC	53.8%	59.3%	80.8%	75.7%	81.9%	**69.3%**
PRE	65.4%	76.1%	80.8%	82.4%	86.8%	**79.3%**

COR = number of correct slot fillers generated by the system
INC = number of incorrect slot fillers generated by the system
PAR = number of partially correct slot fillers generated by the system
POS = number of slot fillers according to the key target template
ACT = number of slot fillers generated by the system
REC = (COR + PAR/2)/POS
PRE = (COR + PAR/2)/ACT

5 Conclusions and Future Work

In this paper we presented a method and its prototypical implementation to provide knowledge modellers of clinical practice guidelines with information about control and data flow, time dimensions, and background information. We identify this information using NLP methods based on both syntactic and semantic information. The majority of the defined information extraction rules are based on semantic relations we derived from the UMLS Semantic Network [11]. By providing this additional information the knowledge modellers see what parts they have to model and what kind of models they have to use.

Our next steps will be the further development of the extraction methods by applying the methods on documents from other clinical specialties. Thereby, we can improve and expand our rule base. Furthermore, we will try to extract more detailed information to automatically generate a MHB model. The correct interpretation of the control dimension will be the major challenge. We have to discover whether there is a single action, a decomposition of an action, or a selection of one or more alternative actions. This support can promote the application of computer-interpretable CPGs.

Acknowledgement. We'd like to thank Andreas Seyfang for his MHB support and we'd like to thank the anonymous reviewers for their valuable feedback. This work is partially supported by "Fonds zur Förderung der wissenschaftlichen Forschung FWF" (Austrian Science Fund), grants L290-N04 and TRP71-N23, and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n º 2161.

References

1. Field, M.J., Lohr, K.N., eds.: Clinical Practice Guidelines: Directions for a New Program. National Academies Press, Institute of Medicine, Washington DC (1990)

2. Quaglini, S., Ciccarese, P., Micieli, G., Cavallini, A.: Non-compliance with guidelines: Motivations and consequences in a case study. In Kaiser, K., Miksch, S., Tu, S.W., eds.: Computer-based Support for Clinical Guidelines and Protocols. Proceedings of the Symposium on Computerized Guidelines and Protocols (CGP 2004). Volume 101 of Studies in Health Technology and Informatics., Prague, Czech Republic, IOS Press (2004) 75–87
3. Peleg, M., Tu, S.W., Bury, J., Ciccarese, P., Fox, J., Greenes, R.A., Hall, R., Johnson, P.D., Jones, N., Kumar, A., Miksch, S., Quaglini, S., Seyfang, A., Shortliffe, E.H., Stefanelli, M.: Comparing computer-interpretable guideline models: A case-study approach. Journal of the American Medical Informatics Association (JAMIA) **10**(1) (Jan-Feb 2003) 52–68
4. Seyfang, A., Miksch, S., Marcos, M., Wittenberg, J., Polo-Conde, C., Rosenbrand, K.: Bridging the gap between informal and formal guideline representations. In Brewka, G., Coradeschi, S., Perini, A., Traverso, P., eds.: European Conference on Artificial Intelligence (ECAI-2006). Volume 141 of Frontiers in Artificial Intelligence and Applications., Riva del Garda, Italy, IOS Press (2006) 447–451
5. Serban, R., ten Teije, A., van Harmelen, F., Marcos, M., Polo-Conde, C.: Extraction and use of linguistic patterns for modelling medical guidelines. Artificial Intelligence in Medicine **39**(2) (2007) 137–149
6. Serban, R., ten Teije, A.: Exploiting thesauri knowledge in medical guideline formalization. Methods Inf Med **48** (2009) 468–474
7. Kaiser, K., Akkaya, C., Miksch, S.: How can information extraction ease formalizing treatment processes in clinical practice guidelines? A method and its evaluation. Artificial Intelligence in Medicine **39**(2) (2007) 151–163
8. National Collaborating Centre for Women's and Children's Health: Induction of labour. Clinical guideline 70, National Institute for Health and Clinical Excellence (NICE), London (UK) (July 2008)
9. Aronson, A.R.: MetaMap: Mapping text to the UMLS metathesaurus. Technical report, Lister Hill National Center for Biomedical Communications (2006)
10. Klein, D., Manning, C.D.: Fast exact inference with a factored model for natural language parsing. In: Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA, MIT Press (2003) 3–10
11. McCray, A.T.: UMLS Semantic Network. In: Proc. of the 13th Annual Symposium on Computer Applications in Medical Care (SCAMC'89). (1989) 503–507
12. Levy, R., Andrew, G.: Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In Fellbaum, C., Miller, G.A., eds.: Proceedings of the fifth international conference on Language Resources and Evaluation (LREC), Genoa, Italy, ELRA (2006) 2231–2234
13. Votruba, P., Miksch, S., Kosara, R.: Facilitating knowledge maintenance of clinical guidelines and protocols. [15] 57–61
14. Institute for Clinical Systems Improvement (ICSI): Management of labor. Clinical guideline, Institute for Clinical Systems Improvement (ICSI), Bloomington (MN) (May 2009)
15. Fieschi, M., Coiera, E., Li, Y.C.J., eds.: Proceedings from the Medinfo 2004 World Congress on Medical Informatics. In Fieschi, M., Coiera, E., Li, Y.C.J., eds.: Proceedings from the Medinfo 2004 World Congress on Medical Informatics, AMIA, IOS Press (2004)