Supporting Computer-interpretable Guidelines' Modeling by Automatically Classifying Clinical Actions

Anne-Lyse Minard and Katharina Kaiser

Institute of Software Technology and Interactive Systems
Vienna University of Technology
Favoritenstrasse 9-11, 1040 Vienna, Austria
{firstname.lastname}@tuwien.ac.at

Abstract. Modeling computer-interpretable clinical practice guidelines is a complex and tedious task that has been of interest for several attempts to automate parts of this process. When modeling guidelines one of the tasks is to specify common actions in everyday's practical medicine (e.g., drug prescription, observation) in order to link them with clinical information systems (e.g., an order-entry system). In this paper we compare a rule-based and a machine-learning method to classify activities according to the *Clinical Actions Palette* used in the Hybrid-Asbru ontology. We use syntactic and semantic features, such as the Semantic Types of the UMLS to classify the activities. Furthermore, we extend our methods by using 2-step classification and combining machine learning and rule-based approaches. Results show that machine learning performs better than the rule-based method on the classification task. They also show that the 2-step classification method improves the categorization of activities.

Keywords: Clinical Practice Guidelines, Hybrid-Asbru, Common Clinical Actions, Natural Language Processing, Classification

1 Introduction

Clinical practice guidelines (CGPs) are important means to provide state-of-the-art medical care in diagnosis and treatment of patients and therefore improve the quality in health care and reduce costs [9]. Computer-interpretable CPGs (CIGs) have been shown to improve the adherence to these guidelines and support the medical personnel by providing patient-specific recommendations at point of care [23]. In order to enable efficient linking of the CIGs on clinical information systems (e.g., order-entry systems), it is necessary to explicitly represent common clinical actions, such as drug prescriptions or physical examinations. Several approaches have been made to classify such clinical actions, for instance, the Unified Service Action Model (USAM) of HL7 RIM [24], the Action Palette by Essaihi et al. [8], or the *Clinical Actions Palette* used in the *Hybrid-Asbru* [27] CIG formalism.

Modeling CIGs is a complex and tedious task that involves the cooperation of both knowledge engineers and medical experts. Automating parts of the modeling process reduces the workload and information extraction techniques are a valuable means for that (e.g., [13]). In order to automatically model procedural parts of the CPGs for a computerized execution, actions need to be specified according to their highest level of detail.

We address the challenge of classifying clinical activities according to the *Clinical Actions Palette* in a way that enables specification of actions in Hybrid-Asbru. Our techniques can be used for supporting or to some extent replacing extracting actions, which is currently accomplished manually by knowledge engineers together with medical experts. Such an automatic classification can then be integrated in a CIG authoring tool to reduce the workload of the modellers. Automatically generated model fragments always need to be manually validated by human experts. However, this validation should be less laborious than a manual classification.

One difficulty we are faced with is to distinguish between two confusable classes, such as *drug-administration* and *drug-prescription* by using sentence elements. In some cases it is difficult even for a human annotator to assign the correct class. Thus, we propose not only sole rule-based or machine learning methods, but also a two-step classification approach, where these methods can be combined.

This paper is organized as follows. In Section 2 we present the context of our work and we make a brief overview of similar works and techniques. A description of the materials and methods is given in Section 3 and then in Section 4 an evaluation of the proposed methods is presented and discussed. Finally, in Section 5 we present our conclusions.

2 Background and Related Work

Asbru [25] is a formalism that represents CIGs as a hierarchy of time-oriented skeletal plans. However, it does not include explicit constructs for expressing common clinical actions such as drug prescription or physical examination. Although these actions are frequently used in everyday's clinical practice the textual nature of the knowledge role can limit its interpretation by the execution engine (e.g., to extract the precise dose of a drug in a drug-prescription action or the name of the laboratory test in an observation).

Hybrid-Asbru is an extension of Asbru which was expanded to include, amongst others, the *Clinical Actions Palette* to explicitly express common clinical actions such as drug prescriptions or physical examinations.

Currently, the actions palette includes the following actions: (1) anamnesis – used to specify querying patients for relevant history; (2) physical-examination – used to specify the performance of various physical examinations to the patient (e.g., measuring heart rate); (3) observation – used to specify an observation like an order of a laboratory test (e.g., WBC count); (4) procedure – used to specify some clinical procedure by a clinician; (5) drug-administration – used to specify the administration of a drug and its details (e.g., route) to a patient by a care-provider; (6) drug-prescription – used to specify a prescription of a drug and its details (e.g., dose) to a patient by a clinician; (7) referral – indicates a referral of a patient to a specialist in a particular medical domain (e.g., endocrinologist); and (8) notification – used to specify advising or educating a patient.

We are focusing on labeling activities according to this *Clinical Actions Palette*. We consider this task as a multi-class classification task where the aim is to categorize a segment of a sentence that describes an action into one of the 8 classes. We work on the segment-level due to the fact that a sentence can contain multiple activities of different types.

Some Natural Language Processing tasks are based on sentence classification, such as text structuring, opinion mining, or sentiment analysis (see [22] for a summary of systems developed for the i2b2 challenge 2011 on sentiment analysis of suicide notes). Khoo et al. [16] evaluated the performance of three classification algorithms (Naive Bayes, Decision Trees, and SVM) for sentence classification in e-mails. By using bag-of-words features, they showed that SVM outperforms the other classification algorithms. McKnight and Srinivasan [21] structured MedLINE abstracts in *Introduction*, *Method*, *Result*, and *Conclusion* sections. They chose SVM and linear classifiers using a large corpus with structured abstracts. There are also approaches using Conditional Random Fields (CRF), as they enable sequential modeling (i.e., taking into account the labeling of adjacent instances and features) [17, 5]. In most cases, the set of features contains bag-of-words features, UMLS concepts or semantic types, features related to the position of the sentence, and sequential features (features from adjacent sentences). In our task there are no discourse dependencies between activities, so the use of CRF is not relevant. To deal with close classes, Chung and Coiera [5] proposed a two-step classification method. From five classes, two of them were very close. In the first step, they gathered the two close classes and classified sentences in the resulting four classes. Using a SVM classifier they learned then to distinguish between the two close classes. This two-step method obtained better results than the initial 5-class classification.

Few works were done on sentence classification in CPGs, for example to detect sentences that contain activities [14] or to classify some kind of activities depending whether they contain a clinical rule, a treatment recommendation, etc. [26]. Bouffier and Poibeau [3] developed a set of rules to detect activity and condition segments in French CPGs. To the best of our knowledge, there are no works which aim to classify activities in CPGs.

3 Materials and Methods

Classifying activities involves representing them by semantic, lexical, syntactic, etc. information in order to find similarities between activities and other activities already categorized. We present in this section our corpus, how we represent information, and the methods employed for the classification.

3.1 Corpus

In order to be able to develop and test our method we built a corpus consisting of eight CPGs. They cover different specialities, such as Cardiology (3), Endocrinology (2), Oncology (2), Pulmonology (1), and were developed by six different institutions (i.e., NICE[1], ACOG[2], CBO[3], SIGN[4], ADA[5], AHA[6]). The guidelines were selected to show whether we can develop a reliable method that is applicable on different medical specialities having varied types of activities as well as on similar guideline topics from different institutions to have a variation in the document structure and language.

We work on activity classification and not on activity extraction, so we use semi-structured texts as input, i.e., texts have been manually annotated with *activity* markups. For the development and the evaluation of our methods, we manually annotated activities in the eight guidelines and classified them according to the *Clinical Actions Palette*. Our corpus contained 348 annotated activities. In Table 1, we indicate the number of activities for each activity type.

In sentence (1) we give an example of a sentence containing an activity, which is assigned the type *drug-prescription*. Sentence (2) shows an example of a recommendation indicating that an action must not be activated. This kind of activity has not been annotated because only the activities that can be executed are interesting for the modeling of the CPGs.

(1) $^{\text{COND}}$ In adult patients with ABPA, $^{\text{ACT}}$a four month trial of itraconazole should be considered.

[1] National Institute for Health and Clinical Excellence; http://www.nice.org.uk

[2] The American Congress of Obstetricians and Gynecologists; http://www.acog.org

[3] Dutch Institute for Healthcare Improvement; http://www.cbo.nl

[4] Scottish Intercollegiate Guidelines Network; http://www.sign.ac.uk

[5] American Diabetes Association; http://www.diabetes.org

[6] American Heart Association; http://www.heart.org

Table 1: Number of activities of each type in the corpus.

Activity Type	**Number**
anamnesis	6
drug-administration	42
drug-prescription	99
notification	46
observation	41
physical-examination	37
procedure	52
referral	25
total	348

(2) [$^{\text{COND}}$ In patients with decompensated HF and AF], [$^{\text{ACT-NOT}}$intravenous administration of a nondihydropyridine calcium channel antagonist [$^{\text{EFFECT}}$may exacerbate hemodynamic compromise] and is not recommended].

3.2 Text Pre-processing

Before the classification task starts, we preprocess the text to obtain semantic and syntactic information from sentences. We used GATE [7], an open source free software for text processing, which provides a set of text engineering tools from which we used the ANNIE tokenizer, sentence splitter, and gazetteer [6], the openNLP[7] POS tagger and chunker, and the MetaMap [2] plugin. In addition, we also developed some handwritten extraction rules and implemented them with JAPE[8]. Figure 1 presents the architecture of the text pre-processing system. The description of some of these modules is following.

GATE's MetaMap plugin maps text with concepts of the UMLS Metathesaurus [19]. We extended it by detecting acronyms and annotating them with their according UMLS concepts. Each UMLS concept is also assigned its *semantic group*, a coarser classification of the semantic types, defined by [20].

We then analyzed two CPGs (Atrial Fibrillation [10] and Gestational Diabetes [1]) according to their activities and the classes they are assigned to. Special emphasis was put on verbs and on other trigger words that could be used to identify the type of activity. We used the ANNIE gazetteer to annotate the verbs and the trigger words in the documents. A gazetteer consists of lists of entities

[7] http://opennlp.apache.org

[8] JAPE (Java Annotation Patterns Engine) provides finite state transduction over annotations based on regular expressions

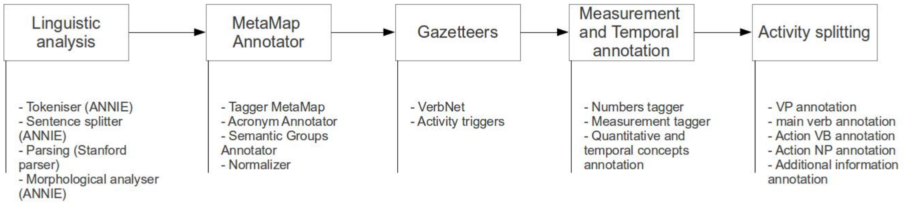

Fig. 1: Pipeline for pre-processing CPGs.

(or more generally of words) which are used to find occurrences of these entities in text. We choose that the matching is done on word lemmas [9].

We used VerbNet [18] to classify verbs according to the VerbNet classes. There are 101 top-level classes and 270 first-level classes that include verbs that are syntactically and semantically close. From the top-level classes we manually selected 15 classes that deemed relevant for identifying different classes of activities according to our analysis (see Table 2 for examples). We have kept verbs that express an activity in our domain, such as verbs of the "removing" and "measure" classes. But we have also retained classes of verbs used to make a recommendation, such as the "communication" class.

Table 2: Examples of verbs in relevant VerbNet classes (in the second column, the first number is the number of the class, and the second one the number of the sub-class).

Top-level classes	**First-level classes**	**Verbs**
Verbs of Assessment	assessment-34.1	analyze, evaluate, review
	estimate-34.2	approximate, count, assess
Indicate verbs	indicate-78	imply, predict, expose

We also observed that there are common words in each kind of activity. For example, in the *notification* activity there are words such as "advice", "inform", etc. We built a list with these activity triggers and used a gazetteer to annotate them in the corpus. These trigger words are categorized in four classes: *trigger_inform*, *trigger_treat*, *trigger_examine*, and *trigger_refer*.

[9] The lemma of a word is its canonical form. For example, the lemma of a noun will correspond to its singular form (the lemma of "symptoms" is "symptom") and the lemma of a verb will often correspond to its infinitive form (the lemma of "achieved" is "achieve").

Furthermore, we used plugins for tagging numbers and measurements as well as quantitative and temporal concepts [11]. With the latter we are able to identify concepts such as age, duration, frequency, and measurement.

3.3 Feature representation

Using lexical features means to be domain-dependent and to have a well-represented training corpus. For example, in diabetes CPGs the UMLS concept "hypoglycaemic therapy" will be a marker for a *drug-prescription* activity, but not in oncology CPGs. We chose to use mainly semantic and syntactic features to obtain a classification model more general and less guideline-dependent. We defined the following features that are extracted from the activity sentence segment:

- Semantic types and semantic groups of UMLS concepts;
- VerbNet classes of verbs;
- The presence of a measurement indication;
- The main verb of the sentence, if it is in the activity sentence segment (i.e., the root node in the dependency graph);
- Triggers of the activity (word and class).

The features extracted, as described in the previous subsection, also contain some noise or are even missing. For example, MetaMap does not recognize all concepts in the text (e.g., because of their particular form, abbreviations, etc.) or assigns wrong semantic types. Also POS tagging or chunking are sources of errors. We did not evaluate the feature extraction process in general, but tried to optimize its output [15] and chose to deal with noise and silence of features.

The features are binary-features that are set to either 0 or 1, depending on whether they are present in the activity sentence segment or not. In total, there are 157 features.

3.4 Rule-based method

We manually developed extraction rules for each of the eight classes. The rules are based on the features described above. Thereby, features are combined and can also be explicitly excluded for a certain activity type. Table 3 shows the number of rules for each activity type and some examples. The example for *drug-administration* means that if a UMLS concept of semantic type "Spatial concept" (spco), a UMLS concept of semantic group "Chemicals and drugs" (CHEM), and the trigger word "administration" appears in the activity clause, it is assigned the "drug-administration" class.

Table 3: Number of rules for each activity type and some examples. Weights of the rules are indicated in brackets.

Activity type	**# rules**	**Example**
anamnesis	1	
drug-administration	6	`IF ST=spco and SG=CHEM and tg=administration` (0.8)
drug-prescription	10	
notification	3	`IF tg_class!=trigger_refer and tg=advise` (0.3)
observation	8	`IF ST=lbtr and tg=check` (0.6)
physical-examination	12	
procedure	8	
referral	5	

Legend: ST=semantic type (w=0.3), SG=semantic group (w=0.2), tg=trigger word (w=0.3), tg_class=trigger class (w=0.2), lbtr= Laboratory or Test results

Each feature in the rule has a weight used to select the correct class in case of multi-label instance. For example, an activity can be classified both in “drug-administration” and “procedure” classes, the matched rule with the higher weight (i.e., the sum of the weights of each feature) will be selected. The features are weighted differently. For instance, the weight for a semantic type feature will be 0.3 while the weight of a VerbNet feature will be set at 0.2. The weights were adjusted by applying rules on the development corpus. The weight for an absent feature (i.e. a feature included in the rule that must be absent of the sentence segment) is null.

The rule base was developed using a molecular approach: we started with developing a set of highly reliable rules and gradually extended our rule base to cover also less frequent patterns. Here, we had also to take care to avoid over-generation of rules by concurrently optimizing recall and precision. In this way, completeness of the rules is not achieved, but the rule set is optimized with regard to precision and recall to also work on new and unseen input.

3.5 Machine-learning method

Next, we used a supervised machine learning approach which uses the features described in subsection 3.3. Supervised machine learning is a technique that automatically learns a model to classify data from a reference corpus in which data has already been classified. The model can then be applied on new data. We have tested different classifiers through the WEKA suite [12], and obtained the best results with a SVM (Support Vector Machine) based classifier called LibSVM [4]. Moreover, SVM-based classifiers are often used for classification tasks in Natural Language Processing domain and is given good results. The SVM method uses an input vectorial representation of data and functions for finding the optimal separation among data. It supports multi-class classification and uses a “one-against-one” approach, i.e., a model is built for each pair of classes and a vote on decision values allows one to obtain one label for each instance. We used LibSVM with a linear kernel and the default parameters.

3.6 Combination of classifiers and rules

In order to improve our “single-step” methods we proposed further methods using combinations of the methods mentioned above (see Figure 2 for the alternative approaches). During our experiments we observed that some classes are confusable. We proposed 2-step methods to improve the classification in these confusable classes: first instances are classified in upper-level classes and then for each upper-level class, they are classified in sub-classes. By combining close classes together, we are able to minimize classification errors in the first step and specialize our classification in the subsequent step.

For example, we observed that the “drug-prescription” (see sentence (3) for an example) and “drug-administration” (see sentence (4)) classes are confusable (e.g. they share semantic properties, such as the words used for expressing an action or the semantic types of the main concepts).

(3) **A single oral bolus dose of propafenone or flecainide (“pill-in-the-pocket”) can be administered** to terminate persistent AF outside the hospital [...]

(4) **Unfractionated heparin may be administered either by continuous intravenous infusion in a dose sufficient** [...]

Thus, we combined the close classes together into 3 upper-level classes *treatment or procedure* (“drug-administration”, “drug-prescription”, and “procedure”), *examination* (“observation” and “physical-examination”), and *other activities* (“referral”, “anamnesis”, and “notification”). So in a first step, activities are classified in these upper-level classes and in a second step, they are then classified in the final activity class. We developed two different methods for this two-step classification:

SVM-SVM classification. We used a SVM classifier to learn to classify activities in the 3 upper-level classes. Then three classifiers were trained to distinguish the sub-classes. The different classifiers use different features. For example, features which represent verbs are more useful for the class “referral” and “notification” than for classes “drug-administration” and “drug-prescription”, and the measurement features are useful for the upper-level classification rather than for distinguishing between “drug-administration” and “drug-prescription”.

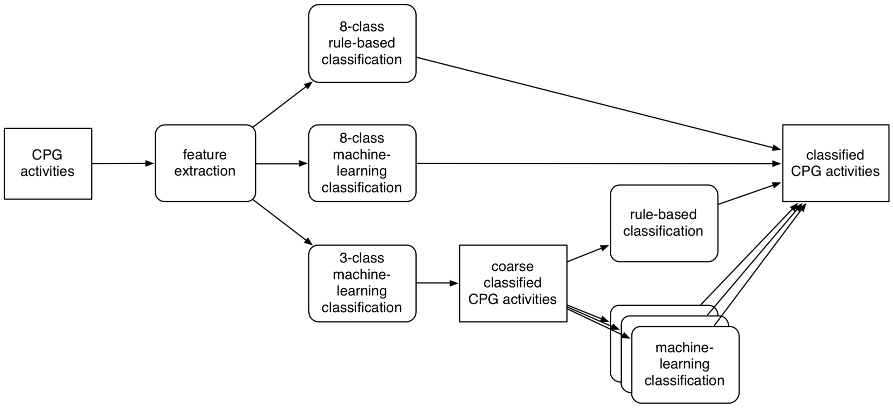
8-class rule-based classification
CPG activities
feature extraction
8-class machine-learning classification
classified CPG activities
rule-based classification
3-class machine-learning classification
coarse classified CPG activities
machine-learning classification

Fig. 2: Classification approaches. After extracting features we applied a rule-based approach (see top) or a SVM-based machine learning approach (below) for classifying activities in one step. At the bottom the 2-step approaches are represented with the SVM-based coarse classification in the first step and then either a rule-based classification or three different SVM-based classifiers.

SVM-rule-based classification. This method uses the same classifier for the upper-level classes than the previous method. For distinguishing among the sub-classes, the rule-based method is used. Rules described in subsection 3.4 are applied on the instances classified in the upper-level classes. For example, on the instances classified in the *examination* upper-level class we apply rules of the "observation" and "physical-examination" sub-classes. Thus, in the second step less rules need to be applied on one activity, which reduces the error rate.

4 Evaluation

For the evaluation of our method, we chose a cross-learning evaluation, i.e., the classifier is trained with 5 CPGs and the two CPGs used for the development and tested on the remaining CPG. By this way we evaluated the system on 6 CPGs using a different training set each time. For the evaluation we employ the classic measures: recall (1), precision (2), F-measure (3), and accuracy (4). The upper-bound of these 4 measures is 1.00. An F-measure of 1.00 means that all the instances which must be classified are classified and that all the instances classified are correctly classified. A perfect accuracy means that the system has correctly classified all the positive instances. The F-measure takes into account positive and negative instances, whereas the accuracy evaluates only the classification of positive instances.

$$Recall = \frac{\text{number of activities correctly classified}}{\text{number of activities to classify}} \tag{1}$$

$$Precision = \frac{\text{number of activities correctly classified}}{\text{number of classified activities}} \tag{2}$$

$$F\text{-}measure = \frac{2 \cdot (\text{precision} \cdot \text{recall})}{(\text{precision} + \text{recall})} \tag{3}$$

$$Accuracy = \frac{\text{total number of activities correctly classified}}{\text{total number of activities}} \tag{4}$$

We also use macro-recall, macro-precision, and macro-F-measure, which correspond to the average of recall, precision, and F-measure respectively of each class.

In Table 4 the results obtained from both the machine learning method and the rule-based method are presented. In the first part, the F-measure obtained for each class is shown for both methods. Then macro-recall, macro-precision, macro-F-measure, and accuracy is given for both methods on the 6 corpora. Bold numbers in macro-F-measure and accuracy indicate the better performing method comparing machine learning and rule-based methods.

Table 4: Results of the 8-class classification with machine learning (ML) and rule based (RB) methods.

	Diabetes type II (ADA)		Pre-eclampsia (ACOG)		Asthma (SIGN)		Breast Cancer (CBO)		Chronic HF (NICE)		Breast Cancer (NICE)	
# activities	33		24		42		7		42		47	
						F-measure						
	ML	**RB**	**ML**	**RB**	**ML**	**RB**	**ML**	**RB**	**ML**	**RB**	**ML**	**RB**
observation	0.29	0.40	0.40	0.00	0.67	0.75	-	-	0.33	0.00	0.22	0.00
physical-examination	0.60	0.67	0.75	0.67	-	-	-	-	0.75	0.36	-	-
referral	0.00	0.00	-	-	1.00	1.00	-	-	0.17	0.53	0.67	0.00
anamnesis	-	-	-	-	0.00	0.00	-	-	-	-	0.00	0.00
notification	0.33	0.40	-	-	0.00	0.00	0.80	0.00	0.18	0.00	0.63	0.00
drug-administration	-	-	0.00	0.67	0.00	0.00	1.00	1.00	-	-	0.00	0.00
drug-prescription	0.50	0.00	0.94	0.57	0.78	0.72	1.00	0.67	0.87	0.75	0.31	0.15
procedure	0.38	0.14	0.55	0.60	0.00	0.44	1.00	0.50	0.20	0.31	0.48	0.30
Macro-recall	0.45	0.41	0.58	0.48	0.39	0.54	0.92	0.50	0.52	0.31	0.47	0.05
Macro-precision	0.40	0.24	0.50	0.58	0.43	0.45	1.00	0.63	0.48	0.40	0.43	0.20
Macro-F-measure	**0.35**	0.27	**0.53**	0.50	0.41	**0.49**	**0.95**	0.54	**0.36**	0.28	**0.33**	0.06
Accuracy	**0.39**	0.33	**0.67**	0.54	0.60	**0.62**	**0.86**	0.43	**0.50**	0.48	**0.36**	0.11

The machine learning method performs better for the classification of activities in 5 of the 6 CPGs. In most cases the classification in the *drug-prescription* and *physical-examination* is good whereas for the other classes the results depend on the corpus. No activities are categorized in the *anamnesis* class neither by the classifier nor by the rules because there are not enough examples of this class in the training corpus (six instances overall). The overall accuracy[10] on the 6 corpora by using the machine learning method is 0.50 and 0.40 with the rule based method. So with the machine learning method half of the activities are correctly classified.

Table 5 shows the results of the 2-step classification methods. In the upper part F-measures of the first step for each of the 3 upper-level classes are given, whereas in the lower part we present the final results obtained after the second classification by using machine learning (ML) or rules (RB). Bold numbers in macro F-measure and accuracy show for which corpus and with which method the performance are better with the 2-step classification compared to the 1-step classification (Table 4).

Table 5: Results of the 2-step classification. In bold the improvements of the 2-step method in comparison to the 1-step method are presented.

	Diabetes type II (ADA)		Pre-eclampsia (ACOG)		Asthma (SIGN)		Breast Cancer (CBO)		Chronic HF (NICE)		Breast Cancer (NICE)	
First-step						**F-measure**						
Treatment/procedure	0.76		0.90		0.88		1.00		0.81		0.81	
Examination	0.80		0.82		0.73		-		0.76		0.67	
Other activities	0.60		-		0.80		1.00		0.85		0.67	
Macro-F-measure	0.72		0.86		0.80		1.00		0.81		0.71	
Second-step	**ML**	**RB**	**ML**	**RB**	**ML**	**RB**	**ML**	**RB**	**ML**	**RB**	**ML**	**RB**
Macro-recall	0.49	0.27	0.61	0.52	0.40	0.53	0.92	0.88	0.55	0.31	0.50	0.05
Macro-precision	0.41	0.17	0.53	0.58	0.43	0.42	1.00	0.83	0.49	0.33	0.41	0.22
Macro-F-measure	**0.39**	0.18	**0.56**	0.51	0.41	0.45	0.95	0.79	**0.40**	0.29	**0.38**	0.08
Accuracy	**0.45**	0.24	**0.71**	0.54	0.60	0.60	0.86	0.71	**0.52**	0.45	**0.40**	0.13

The results for the first-step classification are quite promising. We have an overall F-measure[11] of about 0.82 and an overall accuracy of 0.80 (so 80% of the activities are correctly classified). For the second step we obtain slightly better results than for the 1-step methods of about 0.54 and 0.39 accuracy for the SVM classifier and the rule-based method, respectively. By using rules in the second

[10] The overall accuracy is the sum of all the activities correctly classified through the 6 corpora divided by the number of activities.

[11] The overall F-measure is the average of F-measures obtained from each corpus.

step the classification is still lower than by using only the SVM classifier (in the 1-step method or the 2-step one).

Errors analysis. Analyzing our results we observed different types of errors:

- Features representing an activity were not present in the training corpus: in sentence (5), the verb *discuss* has not been learned from the corpus.
- Wrong or imprecise UMLS semantic type: in example (6), *intramuscular* is linked to the general semantic type *Functional Concept* (rather than to *Spatial Concept*).
- External knowledge is needed to distinguish between two close classes, such as *drug-prescription* and *drug-administration*, e.g., to know if a drug must allways be administered by a physician or could be prescribed.
- Activity annotations may also include other knowledge roles, such as conditions (see sentence (7)), other activities, effects, intentions, etc. This results in a lot of noise in the extracted features.

(5) Their risks and benefits should be **discussed** with the patient and their side-effects carefully monitored.

(6) Consider whether **intramuscular (IM)** hydrocortisone is required.

(7) [...] consider referring **patients with inadequately controlled asthma, especially children**, to specialist care.

To resolve these errors improvements have to be made. First we must increase the size of the training corpus to have a better representation of activity expressions. Then the completion of the trigger list by adding more synonyms could be beneficial to offset the size of the training corpus. Moreover, relating our system to external knowledge like a domain-ontology can enable the system to distinguish between close classes. Finally, a normalization of the sentence to have only information relevant for the activity represented might be beneficial.

5 Conclusion

We have presented a comparison of methods to categorize activities in CPGs according to the *Clinical Actions Palette*. We show that a 2-step method using SVM classifiers is better than a 1-step classification approach using rules or machine learning. Our aim was to deal with a small training corpus by using mainly non-lexical features and also to find a way to classify in confusable classes. For the second issue, involving external resources will be necessary, because the use of rules or a 2-step classification method is not sufficient.

Such a classification of clinical activities can be integrated into a CIG authoring system, but still requires manual validation by human experts. However, an automatic classification will reduce the workload and its validation will still be less burden than a completely manual modeling. Furthermore, in a next step elements and attributes specifying the activity and required for the modeling can be automatically extracted and larger guideline fragments can be generated automatically.

In the future, we plan to extract relevant information from activity segments (e.g., removing some adverbials). We will also work on the detection of relations between activities (e.g., to identify that a drug-prescription activity "Inhaled steroids should be considered for patients with ..." is linked to an adjustment of dose "... and increase the dose of inhaled steroid to 800 mcg/day (adults) ...". And finally, it will be interesting to work on the detection of abort and complete conditions, because these might be formulated similar to activities ("If there is no response to inhaled long-acting beta2 agonist, stop the LABA ...").

Acknowledgements. This research was carried out as part of project no. TRP71-N23 funded by the Austrian Science Fund (FWF) and of the MobiGuide project partially funded by the European Commission under the 7th Framework Program, grant #287811.

References

1. American Diabetes Association: Standards of medical care in diabetes–2011. Diabetes Care 34(Suppl. 1), S11–S61 (2011)
2. Aronson, A.R., Lang, F.M.: An overview of metamap: historical perspective and recent advances. J Am Med Inform Assoc 17, 229–236 (2010)
3. Bouffier, A., Poibeau, T.: Analyzing the Scope of Conditions in Texts: A Discourse-Based Approach. In: Proceedings of the 11th Conference of the Pacific Association for Computational Linguistics. Sapporo, France (2009)
4. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011)
5. Chung, G.Y., Coiera, E.: A study of structured clinical abstracts and the semantic classification of sentences. In: Proc. of the BioNLP Workshop 2007: Biological, Translational, and Clinical Language Processing (BioNLP '07). Association for Computational Linguistics (ACL), Stroudsburg, PA, USA (2007)
6. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of the 40th Anniversary Meeting of the ACL (ACL'02) (2002)
7. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M.A., Saggion, H., Petrak, J., Li, Y., Peters, W.: Text Processing with GATE (Version 6) (2011)
8. Essaihi, A., Michel, G., Shiffman, R.N.: Comprehensive categorization of guideline recommendations: Creating an action palette for implementers. In: AMIA 2003 Symposium Proceedings. pp. 220–224. AMIA (2003)

9. Field, M.J., Lohr, K.N. (eds.): Clinical Practice Guidelines: Directions for a New Program. National Academies Press, Institute of Medicine, Washington DC (1990)
10. Fuster, V., Rydén, L.E., Cannom, D.S., et al.: 2011 ACCF/AHA/HRS Focused Updates Incorporated Into the ACC/AHA/ESC 2006 Guidelines for the Management of Patients With Atrial Fibrillation: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation 123(10), e269–e367 (2011)
11. Gooch, P.: A modular, open-source information extraction framework for identifying clinical concepts and processes of care in clinical narratives. Ph.D. thesis, Centre for Health Informatics, School of Informatics, City University London (2012)
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. 11(1), 10–18 (2009)
13. 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), 151–163 (2007)
14. Kaiser, K., Seyfang, A., Miksch, S.: Identifying treatment activities for modelling computer-interpretable clinical practice guidelines. In: Knowledge Representation for Health-Care, pp. 115–127. No. 6512 in Lecture Notes in Artificial Intelligence, Springer Verlag (2011)
15. Kang, N., van Mulligen, E.M., Kors, J.A.: Comparing and combining chunkers of biomedical text. Journal of Biomedical Informatics 44(2), 354–360 (2011)
16. Khoo, A., Marom, Y., Albrecht, D.: Experiments with sentence classification. In: Procceedings of the 2006 Australasian Language Technology Workshop (ALTW2006). pp. 18–25 (2006)
17. Kim, S.N., Martinez, D., Cavedon, L., Yencken, L.: Automatic classification of sentences to support evidence based medicine. BMC Bioinformatics 12(Suppl 2), S5 (2011)
18. Kipper, K., Korhonen, A., Ryant, N., Palmer, M.a.: A large-scale classification of English verbs. Language Resources and Evaluation 42(1), 21–40 (2008)
19. Lindberg, D., Humphreys, B.L., McCray, A.T.: The unified medical language system. Methods of Information in Medicine 32(4), 281–291 (1993)
20. McCray, A.: An upper-level ontology for the biomedical domain. Comp Funct Genomics 4(1), 80–4 (2003)
21. McKnight, L., Srinivasan, P.: Categorization of sentence types in medical abstracts. In: Proc. of the AMIA Annual Symposium. pp. 440–444 (2003)
22. Pestian, J.P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K.B., Hurdle, J., Brew, C.: Sentiment analysis of suicide notes: A shared task. Biomedical Informatics Insights 5, 3–16 (01 2012)
23. Quaglini, S.: Compliance with clinical practice guidelines. In: ten Teije, A., Miksch, S., Lucas, P.J. (eds.) Computer-based Medical Guidelines and Protocols: A Primer and Current Trends, Studies in Health Technology and Informatics, vol. 139, chap. 9, pp. 160–179. IOS Press (2008)
24. Schadow, G., Russler, D.C., Mead, C.N., McDonald, C.J.: Integrating medical information and knowledge in the HL7 RIM. In: Proceedings of the AMIA Annual Symposium. pp. 764–748 (January 2000)
25. Shahar, Y., Miksch, S., Johnson, P.: The Asgaard project: A task-specific framework for the application and critiquing of time-oriented clinical guidelines. Artificial Intelligence in Medicine 14, 29–51 (September 1998)
26. Song, M., Kim, S., Park, D., Lee, Y.: A multi-classifier based guideline sentence classification system. Healthc Inform Res 17(4), 224–31 (2011)
27. Young, O., Shahar, Y., Liel, Y., Lunenfeld, E., Bar, G., Shalom, E., Martins, S.B., Vaszar, L.T., Marom, T., Goldstein, M.K.: Runtime application of Hybrid-Asbru clinical guidelines. Journal of Biomedical Informatics 40, 507–526 (2007)