Process Fragment Recognition in Clinical Documents

Camilo Thorne[1], Elena Cardillo[2], Claudio Eccher[2], Marco Montali[1], and Diego Calvanese[1]

[1] Free University of Bozen-Bolzano, 3 Piazza Domenicani, 39100, Italy
{thorne,montali,calvanese}@inf.unibz.it
[2] Fondazione Bruno Kessler, 18 Via Sommarive, 38123, Italy
{cardillo,eccher}@fbk.edu

Abstract. We describe a first experiment on automated activity and relation identification, and more in general, on the automated identification and extraction of computer-interpretable guideline fragments from clinical documents. We rely on clinical entity and relation (activities, actors, artifacts and their relations) recognition techniques and use MetaMap and the UMLS Metathesaurus to provide lexical information. In particular, we study the impact of clinical document syntax and semantics on the precision of activity and temporal relation recognition.

Keywords. Clinical entity and relation recognition, UMLS Metathesaurus, natural language processing, process fragment recognition.

1 Introduction

Clinical practice guidelines are systematically developed documents that specify the activities, resources and personnel required to cure or treat a particular illness or medical condition, see [6]. The necessity to instantiate them into clinic and hospital protocols and workflows has given rise to *computer-interpretable guidelines* (CIGs), see [3], viz., formal representations (constructed typically with process representation languages) of the care process or plan. On the other hand, several natural language processing (NLP) techniques have been developed to fully or partially automate their processing, see [10], until now manual, and therefore both time and cost consuming.

Clinical NLP approaches leverage on a number of clinical and biomedical annotation resources. Above all, they rely on the crucial US National Library of Medicine's Unified Medical Language System (UMLS) Metathesaurus[1], which thanks to the two annotation tools built upon it, MetaMap and SemRel (see [1]), has become the key lexical semantics resource in this domain. The UMLS Metathesaurus is a biomedical lexical resource (similar to WordNet) comprising over 1 million biomedical concepts (identified by a CUI – concept unique identifier) and covering over 5 million terms, which stem from the over 100

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

M. Baldoni et al. (Eds.): AI*IA 2013, LNAI 8249, pp. 227–238, 2013.

incorporated controlled vocabularies, nomenclatures and classification systems integrated in it. Concepts in UMLS are structured in a semantic network (or ontology) composed of 150 categories (called "concept types") and 54 semantic relationships.

In this paper we describe some preliminary experiments on how to apply fully-supervised clinical entity recognition techniques inspired by [2] to recognize CIG fragments in medical documents, by leveraging on UMLS concept type and relation annotations. The process dimension of CIGs consists of four pillars: activities, resources, actors, and control flows. We focus on activities, the main building block of CIGs, and their basic temporal relations (before/after). To a lesser extent, we focus also on resources, actors and causal relations. We rely on MetaMap annotations and evaluate our techniques over a small UMLS-annotated clinical corpus, the SemRep corpus. We focus in particular on the issue of feature extraction and selection, to assess which features, be them semantic or (morpho)syntactic, are reasonably good predictors for activity and temporal relation recognition.

This paper is structured as follows. In Section 2 we give an overview of related work on clinical NLP, data mining and unsupervised CIG extraction methodologies. In Section 3 we provide the formal background of CIG fragment recognition. In Section 4 we describe the SemRep corpus, our experimental setting and the goals pursued. In Section 5 we describe and discuss the results of our experiments. Finally, we sum up our conclusions in Section 6, and point out how we intend in the future to study further automated CIG fragment extraction.

2 Related Work

The UMLS Metathesausus has been used for clinical text or data mining purposes in many projects. In the medical domain, early works are MedLEE [8] or MedSyndicate [9]. Also MedIE [21] and SeReMeD [4] apply semantic tagging using the UMLS, but their application is limited to processing radiology reports. Meystre and Haug propose a NLP-based system to extract medical problems from electronic patient records [14]. Clinical NLP frameworks such as cTAKES, proposed by [17], use it for document indexing and retrieval. It has also given rise to automated semi- and fully-supervised annotation techniques and resources: It has inspired the annotation formats used to build clinical annotated corpora such as the CLEF corpus from [16]. Furthermore, as Ben Abacha and Zwiegenbaum in [2] show, it is a key tool for automated clinical entity recognition and for clinical relation recognition and extraction.

The much more complex domain of clinical guidelines and CIGs has been tackled on the other hand, using unsupervised techniques. Serban et al. [18] defined a set of linguistic patterns, which can be used to formally represent the knowledge about medical actions contained in guideline text. By semantic tagging of guidelines, these patterns were identified in the document. After combining them with medical domain knowledge this enables an easier formalization and maintenance of guideline models. A similar approach was pursued by Kaiser et al. [10],

who defined syntactic and semantic patterns that are used to develop extraction rules to identify and extract actions and processes out of guidelines. The patterns were based on the UMLS Semantic Network and its semantic relations. But unsupervised techniques are problematic, because the expert knowledge needed to hand-craft such rules is typically much scarce than training corpora.

We believe that UMLS-based supervised clinical entity recognition and annotation techniques and CIG extraction techniques can be successfully combined by following recent advances from the business process modeling community. Friederich et al. [7] and Di Ciccio and Metella [5] have experimented with pipelines combining lexical resources plus supervised (e.g., parsing, entity recognition) and unsupervised (e.g., control flow patterns) NLP techniques to extract, resp., model fragments from formal requirement documents, and process fragments from emails, both with reasonable amounts of success. Following this intuition we experiment in this paper with a UMLS-driven supervised approach that aims at recognizing activities and temporal relations in clinical documents.

3 Process Fragment Recognition

CIG Fragments. There are several ways to formally characterize CIGs. For convenience, we use terminology coming from the Business Process Modeling and Notation (BPMN) standard (see [13]). A CIG is a complex object constituted by the following basic components:

- static components: *(i) activities* (e.g., providing advice, controlling blood glucose levels), representing units of execution in the process; *(ii)* activity agents, viz., the *actors* (e.g., doctors, nurses, patients); *(iii)* artifacts and data used or consumed by activities or *resources* (e.g., metmorfin);
- dynamic components: *(iv) control flows* (e.g., sequence and "if... then... else" control structures) that specify the acceptable orderings among activities.

To extract CIG components the parse trees and MetaMap annotations of guidelines must be mined. Firstly, noun phrases (**NP**s) and verbs must be identified in the parse or constituency trees. At a second step linguistic and domain knowledge in the form of semantic annotations and constituency relations must be considered. In Figure 1 (top) the reader can see a guideline fragment (recommendation 1.4.1 of the NICE diabetes-2 guideline[2]) with its entities highlighted and their candidate annotations. Activities are not only referred to by verbs but also by nouns and noun phrases: to correctly extract the "deep" intended CIG fragment (see Figure 1, bottom left) it is necessary to "filter out" the two wrong "clinical attribute" annotations. Moreover, we need to realize that the verb "continue" introduces a third activity, and rely on syntactic structure to properly order the activities.

We would like to know if the CIG fragment extraction, and in particular the recognition of the activities and control flows intended in clinical documents

[2] `http://www.nice.org.uk/nicemedia/pdf/CG66NICEGuideline.pdf`

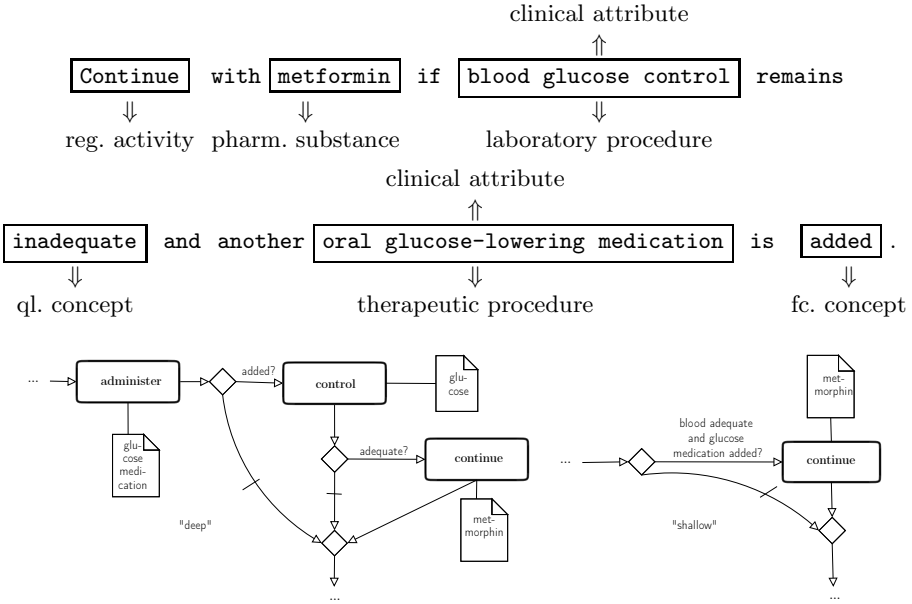


Fig. 1. Top: MetaMap UMLS (automated) annotations of the NICE diabetes guideline fragment; boxes surround entities, annotations are MetaMap's. **Bottom:** Two candidate CIG fragments (represented in BPMN): to the left, the intended "deep" CIG, to the right a "shallow" CIG. Control flows (diamonds) specify the acceptable orderings of the activities (rounded rectangles); activities consume resources (folded-corner rectangles).

can be pursued via supervised clinical entity and relation recognition using an UMLS-annotated corpus for training.

Clinical Entity Recognition. To identify activities and relations in clinical documents, we need to recognize CIG fragments. Let $\boldsymbol{t}_c = (c_1, \ldots, c_n)^T$ denote a vector of n *entity type labels* drawn from a set $\{c_1, \ldots, c_k\}$ of k clinical entities; or, resp., a vector $\boldsymbol{t}_r = (r_1, \ldots, r_n)^T$ of n *relation labels* drawn from a set $\{r_1, \ldots, r_p\}$ of p clinical relations. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ be a vector of n input *noun phrases* (**NP**s) or *entities* (the n **NP**s of a sentence), or resp., a vector $\boldsymbol{\alpha} = ((\alpha_1, \alpha_1), (\alpha_1, \alpha_2), \ldots, (\alpha_n, \alpha_n))^T$ of $n \times n$ input **NP** pairs or relation *arguments* (the $n \times n$ possible pairs of **NP**s in a sentence). The goal of *clinical entity* or, resp., *relation recognition*, see [2], can be formulated as the task of finding the best scoring vector $\boldsymbol{t}^*_\beta$:

$$\boldsymbol{t}^*_\beta = \arg\max_{\boldsymbol{t}_\beta} \mu(\rho(\boldsymbol{\alpha}, \boldsymbol{t}_\beta)) \tag{1}$$

where: $\beta \in \{c, r\}$; $\mu(\cdot)$ denotes a *recognizer* built using a classification model (e.g., a logistic regression or neural network algorithm); and $\rho(\cdot, \cdot)$ is a *feature extraction* function, that maps $\boldsymbol{t}_\beta$ and $\boldsymbol{\alpha}$ into a high-dimensional space of numeric, categorical or ordinal *features* over which the classifier is defined. We study

this task w.r.t. the set {activity, resource, actor, other} of entity type labels and the set {temporal, causal, other} of relation labels, and consider *supervised* recognizers, viz., recognizers that can be estimated from a training corpus.

4 Experiments

Features. Our experiments focused in understanding the predictive power of syntax and semantics for recognizing both clinical activities and their temporal relations. Thus, we decided to use linguistically "deep" features extracted from constituency parse trees in addition to semantic annotations. Following strategies similar to the work proposed by [20] we used the Stanford parser (see [12]) to extract syntactic features, and MetaMap to harvest clinical entities and relations via the UMLS concept types they subsume (see Table 1, top), and to compute the lexical semantic features.

By mining parse trees we extracted from **NP**s the following syntactic features: **(1)** depth *nest* of nesting; **(2)** position *pos* in the phrase; and **(3)** occurrence *sub* in a subordinated phrase. The lexical semantic features were extracted by computing several measures of label overlap and frequency. We extracted also the following semantic features: **(4)** the (raw) frequency of the **NP** entity type c in the corpus; **(5)** the degree of *annotation overlap* φ_{hd} between the (possibly repeated) labels *labs* collected using MetaMap from all the constituent nouns of a **NP**, and the (possibly repeated) labels of its head noun *labsh*; **(6)** the *relative frequency* φ_{lf} of the **NP** entity type c w.r.t. *labs*; and **(7)** *label overlap* φ_{ls} that takes into account the taxonomic structure of the UMLS Metathesaurus; viz., respectively,

$$\varphi_{hd}=\frac{||labs \Cap labsh||}{||labs||+||labsh||} \qquad \varphi_{lf}=\frac{||labs \Cap \{c\}||}{||labs||} \qquad \varphi_{ls}=\frac{||labs \Cap sub(c)||}{||labs||+||sub(c)||} \tag{2}$$

where $||\cdot||$ and $\Cap$ denote resp. bag cardinality and intersection, and $sub(c)$ is the bag of all the UMLS concept types that the entity type label c subsumes. In all cases a simple Laplace smoothing was later applied to prevent division by zero errors. See Table 1, bottom.

The SemRep Corpus. Since no UMLS annotated guideline corpora are available for research purposes we ran our experiments over the SemRep corpus (see [11]), a small annotated clinical corpus. It consists of 500 clinical excerpts (MedLine/PubMed) and contains 13,948 word tokens, manually annotated by clinicians and domain experts, covering the whole clinical domain. UMLS concept types annotate a total of 827 **NP**s (at an average of 2 per sentence). In addition to this, UMLS relations annotate around 200 **NP** pairs.

The domain of SemRep largely overlaps with that of clinical guidelines. Furthermore, they are similar in syntactic structure. Such syntactic structure can be approximated by observing the distribution of function "process-evoking" words (PEWs)[3]. PEWs are tokens belonging to the following word categories:

[3] For the part-of-speech tagging we relied on a 3-gram tagger, with 2-gram and unigram backoffs, trained over the Brown corpus; the trained tagger had 0.8 accuracy.

Table 1. Top: Entity types and sample UMLS concept types they subsume; relations and sample UMLS relations they subsume. **Bottom:** Features considered.

activity	actor	resource	other
laboratory procedure	organization	pharmacological substance	qualitative concept

temporal	causal	other
precedes	prevents	located_in
coexists_with	produces	part_of

feature F	description	value f
nest	nesting level in tree	integer $\in \mathbb{N}$
pos	position w.r.t. verb	subject, predicate
sub	occurs in clause?	yes, no
freq	freq. of label in corpus	integer $\in \mathbb{N}$
φ_{lf}	relative frequency of label in **NP**	real $\in [0, 1]$
φ_{hd}	head/**NP** overlap	real $\in [0, 1]$
φ_{ls}	label/**NP** overlap	real $\in [0, 1]$
class	**NP** entity type	activity, actor, resource, other
rel	relation	temporal, causal, other

- conjunctions and prepositions: subordinating prepositions and conjunctions, e.g., "if"; coordinating conjunctions, e.g., "and", "or";
- adverbs: base adverbs, e.g., "after"; comparative adverbs, e.g., "later"; superlative adverbs, e.g., "latest"; and adverbial particles, e.g., "go back".

We thus compared to SemRep: **(1)** a subset of the NICE diabetes-2 guideline (therapy recommendations, 7,109 words); **(2)** a subset of the NICE eating disorders guideline[4] (therapy recommendations, 5,078 words); and **(3)** a subset of the NICE schizophrenia guideline[5] (therapy recommendations, 5,367 words). We also tried to assess whether there is a significant bias in clinical documents towards PEWs. To this end we compared SemRep and the guideline corpora to: **(4)** a subset of the Brown corpus[6] (A: press articles, 1,391,708 words); **(5)** Friederich's corpus of business process specifications [7] (3,824 words). See Figure 3. We ran two statistical tests:

1. A t-test with the null hypothesis H_0 that cross-corpora PEW mean relative frequency is $\mu_0 = 0.20$ ($p = 0.01$ significance level). This test showed no (statistically) significant differences in PEW distribution across corpora, where about 17 to 20% of word tokens are PEWs (see Section 3).
2. A χ^2-test of (in)dependence, with the null hypothesis H_0 that PEW relative frequency is correlated to (or depends on) corpus domain ($p = 0.01$ significance level). This test gave no (statistically) significant differences across domains (see Section 3).

Figure 2 seems to indicate that syntax is uniform across domains. This seems to justify at the same time the use of SemRep to experiment with CIG fragment

[4] `http://www.nice.org.uk/nicemedia/live/10932/29218/29218.pdf`
[5] `http://www.nice.org.uk/nicemedia/live/11786/43607/43607.pdf`
[6] `http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml`

recognition techniques, as well as the general syntactic features (see Section 3) we extracted. It also suggests that UMLS annotations are independent from syntax. In the following experiments we will thus try to empirically validate the following claim:

Semantic features and environment are more significant for activity and temporal relation recognition than syntactic features. (3)

Feature Vectors. Our main goal was to study feature performance rather than model performance *per se*; we thus relied on standard classification models from the known Weka[7] data mining framework rather than on more sophisticated models. The performance of simple models might be suboptimal, but will nevertheless exhibit recognizable (cross-classifier) trends relatively to the independent features selected for training and prediction. We extracted three sets of observations for our experiments:

1. a set of **NP** observations: for each **NP** α in SemRep, we extracted the feature vector $(f_1^\alpha, \ldots, f_7^\alpha, c^\alpha)^T$;
2. a set of sentence observations: for each vector $(\alpha_1, \ldots, \alpha_k)^T$ of (manually annotated) **NP**s in a SemRep sentence, we extracted the feature vectors of the form $(f_1^{\alpha_1}, \ldots, f_7^{\alpha_1}, c^{\alpha_1}, \ldots, f_1^{\alpha_k}, \ldots, f_7^{\alpha_k}, c^{\alpha_k})^T$;
3. a set of relation observations: for each vector $(\alpha, \alpha', r)^T$ of UMLS annotated **NP**s and their UMLS relation r, we extracted the feature vectors $(f_1^\alpha, \ldots, f_7^\alpha, c^\alpha, f_1^{\alpha'}, \ldots, f_7^{\alpha'}, c^{\alpha'}, r)^T$.

corpus tests

χ^2	p	df.	t-score	p	df.
1.03	0.36	5	43.13	0.00	2

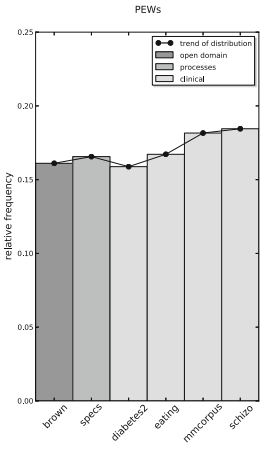


Fig. 2. By "mmcorpus" we mean SemRep

We proceeded to build three parallel sets of training and evaluation observations based on a 2/5 vs. 3/5 split. We considered as feature performance metric, for $\tau \in \{\text{activity}, \text{temporal}\}$, *activity precision* and temporal *relation precision*

$$PR = \frac{|\text{true } \tau\text{s}|}{|\text{true } \tau\text{s}| + |\text{false } \tau\text{s}|}. \tag{4}$$

First Experiment: Feature Significance for Activity Recognition. Our first experiment was designed to measure the significance of each single feature for the activity recognition task. To this end we removed each time a feature F_i from the set $\{F_1, \ldots, F_7\}$ of (syntactic and semantic) *independent* features from Table 1, retrained and measured activity precision w.r.t. $\{F_1, \ldots, F_{i-1}, F_{i+1}, \ldots, F_7\}$.

[7] `http://www.cs.waikato.ac.nz/~ml/weka/`

We considered only **NP** observations. We trained and evaluated the following (Weka) classifiers: logistic classifier, support vector machine, neural network, Bayes classifier, decision tree and a 10-nearest neighbor classifier (baseline).

Second Experiment: Feature Significance for Relation Extraction. Our second experiment was designed to measure the significance of each single feature for the relation recognition task. We proceeded as before, with the difference that we considered an extended set $\{F_1, \ldots, F_8\}$ of 8 independent features including the semantic *class* (i.e., entity type labels) feature, since it arguably describes the *type* of the relation, viz., its domain and range. We considered for this experiment the set of relation observations. We trained and evaluated the following (Weka) classifiers: logistic classifier, neural network, Bayes classifier (baseline) and a decision tree.

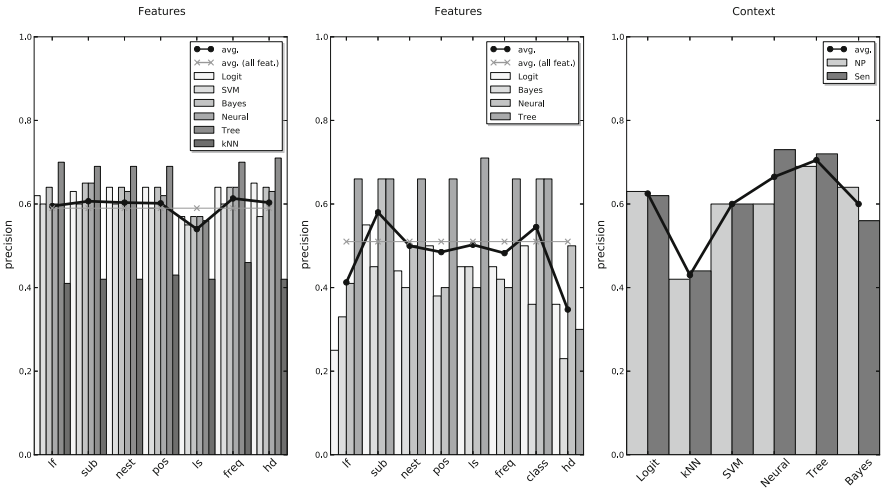
Third Experiment: Context Significance. Our third experiment had the aim of understanding whether sentence context as opposed to **NP** context yields a significant improvement of (average) activity recognition precision. Also, we tried to determine which kind of classifier may perform better. We trained and evaluated the following (Weka) classifiers: logistic classifier, support vector machine, neural network, Bayes classifier, decision tree, and 10-nearest neighbor classifier (baseline) over *(i)* **NP** observations and *(ii)* sentence observations.

5 Evaluation and Discussion

The first experiment (see Figure 3, left), shows a statistically significant drop in cross-classifier average precision when label/**NP** overlap is disregarded, viz., a drop from 0.72 (the precision of the decision tree over the full set on **NP** observations) to 0.56. The removal of syntactic features had little to no effect, and the removal of label relative frequency and head/**NP** overlap gave rise to a slight performance decrease. These results suggest that the (lexical) semantic environment of **NP**s is more relevant (on average) for activity recognition than its syntactic environment.

The second experiment shows a statistically significant drop in cross-classifier average precision w.r.t. the best performing classifier, when head/**NP** overlap and label frequency are disregarded, viz., a drop from 0.66 (the precision of the decision tree over the full set on relation observations) to 0.4 and 0.33 resp.. It also showed a statistically significant improvement when subordination is disregarded (viz., from 0.66 to 0.58). This may suggest also that the (lexical) semantic environment of relations is more relevant (on average) for relation recognition. Interestingly, and contrary to our expectation, disregarding the typing of the relation seemed also to boost performance.

In the third experiment (see Figure 3, right) the classifier that performed better w.r.t. **NP** observations was the decision tree classifier (0.69 precision), likely because of its exploiting better the categorical features (position, subordination). It also matched over sentence observations the best performing classifier, the neural network, which showed a statistically significant improvement (from 0.60 to



experiment one

χ^2	p-value	df.
408242025.00	0.0	6

experiment two

χ^2	p-value	df.
18550249.00	0.0	7

experiment three

χ^2	p-value	df.
18576100.02	0.0	5

Fig. 3. Left: Experiment one: impact of removing a feature on activity precision, by classifier. **Center:** Experiment two: impact of removing a feature on relation precision, by classifier. **Right:** Experiment three: impact of context on activity precision. For the significance tests we considered as null hypothesis H_0 the uniform distribution.

0.73). This experiment seems to indicate overall that sentence environment is an important factor for activity recognition.

While the experiments seem to substantiate claim (3), in some cases (e.g., experiment two) we obtained also results that may seem to infirm it: a relation's argument type should be a relevant feature for relation recognition. The SemRep corpus is a small and sparsely annotated corpus – for, e.g., experiment two we extracted only 200 observations. Indeed, the patterns identified in Figure 3 were noisy and error-prone: the average cross-experiment classifier accuracy was in general low, see Table 2. Such behavior contrary to our expectations is likely due to the small size of the dataset.

On way to estimate how much data we actually need to obtain reasonably accurate predictions is to apply the theory of probably approximately correct (PAC) learning [19]. This theory implies that to learn from categorical data with $1 - \delta$ confidence a decision tree classifier m of bounded depth $\leq k$ and n attributes that is is $1 - \epsilon$ accurate[8], we need to train m over

$$N \geq \frac{1}{\epsilon} \times (\ln \frac{1}{\delta} + \ln |\mathcal{M}|) \tag{5}$$

observations, where $\mathcal{M}$ denotes the space of all decision trees m.

[8] Parameters ϵ and δ denote classification and learning *error*, resp.

Table 2. Precision, recall and accuracy (all features). In gray, the results for the decision tree classifier.

NP obs.

model	PR	RE	AC
logit	0.63	0.71	0.64
SMV	0.60	0.78	0.62
Bayes	0.64	0.64	0.55
neural	0.60	0.76	0.64
tree	0.69	0.75	0.69
10-nn	0.43	0.72	0.42
(avg)	0.60	0.73	0.59

sentence obs.

model	PR	RE	AC
logit	0.62	0.69	0.63
SMV	0.60	0.73	0.63
Bayes	0.56	0.51	0.54
neural	0.73	0.74	0.73
tree	0.71	0.79	0.73
10-nn	0.43	0. 28	0.41
(avg)	0.61	0.62	0.61

relation obs.

model	PR	RE	AC
logit	0.50	0.50	0.67
Bayes	0.45	0.62	0.71
neural	0.44	0.50	0.73
tree	0.66	0.50	0.73
(avg)	0.51	0.54	0.71

We can apply this result to our setting as follows. We restrict attention to the decision tree models and "discretize" our continuous numeric features from Table 1. Consider experiments one and three where we consider $n = 7$ features. Set to $k = 5$ the depth bound (viz., the number of "good" features for activity or temporal relation recognition suggested by our experiments). Then $|\mathcal{M}| \approx O(7^5)$. To learn with $1-\delta \geq 0.95$ confidence a decision tree m with $1-\epsilon \geq 0.8$ accuracy we need to train it, by applying equation (5), on approx. $N \geq 84050$ observations. In other words, to reach reasonably accurate results we would need approx. 100 times more UMLS-annotated **NP**s than the 827 extracted from SemRep.

6 Conclusions and Further Work

We have conducted a preliminary experiment on how to automatically recognize activities and temporal relations using MetaMap and the UMLS Metathesaurus. We used the UMLS-annotated SemRep corpus as our training and evaluation corpus. We focused in the issue of feature selection, seeking to determine if semantics is more relevant than syntax for this task, and hence on feature performance rather than on classifier performance.

Our experiments have shown that in general the lexical semantic environment of an entity is more significant than its syntactic environment for identifying activities. Corpus analysis on SemRep and other clinical and non-clinical corpora showed moreover that the syntax of clinical text is not significantly different both within and across domains. Taking into consideration sentence context gave rise to a slight gain in performance. In all of our experiments the best performing of all the simple annotators used turned out to be the decision tree, better adapted to the categorical features we considered. The small size of the corpus and in particular the small number of relation annotations made our results much less conclusive however regarding temporal relations.

In the future, we plan to consider more powerful techniques, more complex feature sets, and larger corpora to improve our results. Regarding techniques, we intend to use more powerful classification models for NLP such as conditional random fields (CRFs), which can exploit possible dependencies among independent features. Furthermore, such models allow for very complex linguistic

features and context models (based on n-grams) that we did not, for the sake of simplicity and scope, consider in this paper, such as the bag of n-words or n-POSs surrounding an entity, or the n-typed dependencies in which it participates, to name three. We intend also to consider a bigger corpus by integrating SemRep with the i2b2 clinical corpus as suggested by [2]. Finally, we will experiment with temporal relation extraction methods (*à la* TimeML) to tackle CIG control flow extraction. In fact, the current investigation focuses only on before/after temporal relations among tasks, but our final objective is the extraction of complex CIG fragments encompassing also gateways and more elaborated constraints on the process control-flow. Since the nature of the extracted constraints is declarative, we will not only focus on "procedural" specification languages (such as Asbru, Glare, BPMN), but we will also consider, at least as an intermediate format, constraint-based languages such as CigDec [15].

Acknowledgments. The present work has been done within the context of the VeriCliG project supported by a grant from the Free University of Bozen-Bolzano Foundation.

References

1. Aronson, A.R., Lang, F.-M.: An overview of MetaMap: Historical perspective and recent advances. J. of the American Medical Informatics Association 17(3), 229–236 (2010)
2. Ben Abacha, A., Zweigenbaum, P.: Medical entity recognition: A comparison of semantic and statistical methods. In: Proc. of the BioNLP 2011 Work. (2011)
3. De Clercq, P., Kaiser, K., Hasman, A.: Computer interpretable medical guidelines. In: Ten Teije, A., et al. (eds.) Computer-based Medical Guidelines and Protocols: A Primer and Current Trends, ch. 2, pp. 22–43. IOS Press (2008)
4. Denecke, K.: Structuring of and information extraction from medical documents using the UMLS. Methods of Information in Medicine 47(5), 425–434 (2008)
5. di Ciccio, C., Metella, M.: Studies on the discovery of declarative control flows from error-prone data. In: Proc. of the Third International Symposium on Data-Driven Process Discovery and Analysis, SIMPDA 2013 (2013)
6. Field, M.J., Lohr, K.N. (eds.): Clinical Practice Guidelines. Directions for a New Program. National Academy Press (1990)
7. Friedrich, F., Mendling, J., Puhlmann, F.: Process model generation from natural language text. In: Mouratidis, H., Rolland, C. (eds.) CAiSE 2011. LNCS, vol. 6741, pp. 482–496. Springer, Heidelberg (2011)
8. Friedman, C., Hripcsak, G.: Evaluating natural language processors in the clinical domain. In: Proc. of the Conf. on Natural Language and Medical Concept Representation (1997)
9. Hahn, U., Romacker, M., Schulz, S.: MEDSYNDICATE–A natural language system for the extraction of medical information from findings reports. Int. J. of Medical Informatics 67(1-3), 41–52 (2002)
10. Kaiser, K., Akkaya, C., Miksch, S.: How can information extraction ease formalizing of treatment processes in clinical practice guidelines? Artificial Intelligence in Medicine 39(2), 151–163 (2007)

11. Kilicoglu, H., Rosenblat, G., Fiszman, M., Rindfleisch, T.C.: Constructing a semantic predication gold standard from the biomedical literature. BMC Bioinformatics 12(486) (2011)
12. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics, ACL 2003 (2003)
13. Ko, R.K.L., Lee, S.S.G., Lee, E.W.: Business process mangament (BPM) standards: A survey. Business Process Management J. 15(5), 744–791 (2009)
14. Meystre, S., Haug, P.: Natural language processing to extract medical problems from electronic clinical documents. J. of Biomedical Informatics 39(6), 589–599 (2006)
15. Mulyar, N., Pesic, M., van der Aalst, W.M.P., Peleg, M.: Declarative and procedural approaches for modelling clinical guidelines: Addressing flexibility issues. In: ter Hofstede, A.H.M., Benatallah, B., Paik, H.-Y. (eds.) BPM 2007 Workshops. LNCS, vol. 4928, pp. 335–346. Springer, Heidelberg (2008)
16. Roberts, A., Gaizaskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J., Roberts, I., Setzer, A., Tapuria, A., Wheeldin, B.: The CLEF corpus: Semantic annotation of a clinical text. In: Proc. of the AMIA 2007 Annual Symp. (2007)
17. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. J. of the American Medical Informatics Association 17(5), 507–513 (2010)
18. Serban, R., ten Teije, A., van Harmelen, F., Marcos, M., Polo-Conde, C.: Extraction and use of linguistics patterns for modelling medical guidelines. Artificial Intelligence in Medicine 39(2), 137–149 (2007)
19. Valiant, L.G.: A theory of the learnable. Communications of the ACM 27(11), 1134–1142 (1984)
20. Zhou, D., He, Y.: Semantic parsing for biomedical event extraction. In: Proc. of the Ninth Int. Conf. on Computational Semantics, IWCS 2011 (2011)
21. Zhou, X., Han, H., Chankai, I., Prestud, A., Brooks, A.: Approaches to text mining for clinical medical records. In: Proc. of the 2006 ACM Symposium on Applied Computing (2006)