VERICLIG: Extraction and Verification of Clinical Guidelines

Project Report for 2013

Camilo Thorne (I), Marco Montali (I), Diego Calvanese (PI)

Faculty of Computer Science
Free University of Bozen-Bolzano
Piazza Domenicani 3, Bolzano, Italy
{cthorne,montali,calvanese}@inf.unibz.it
http://www.inf.unibz.it/~cathorne/vericlig

March 1, 2014

Abstract

The following document reports on the activities and results of the VERICLIG project during its second and last year (February 2013–January 2014). VERICLIG was funded by a grant from the Free University of Bozen-Bolzano Research Foundation from February 2012 until January 2014. We summarize the research problem studied during this period, namely, extracting workflow components (activities, actors and resources) from clinical documents, the techniques (clinical entity recognition) applied to tackle it and their evaluation. We also provide a list of all the publications, visits, talks and collaborations related to the project or that were financially supported by it.

Contents

1.5.1.2 Emphasise advice on healthy balanced eating that is applicable to the general population when providing advice to people with type 2 diabetes.

1.5.1.3 Continue with metformin if blood glucose control remains inadequate and another oral glucose-lowering medication is added.

Figure 1: An excerpt from the NICE diabetes-2 clinical guideline[3]. Each line describes atomic treatments that combine together into a complex therapy.

1 Introduction

Clinical guidelines are evidence-based documents compiling the best practices for the treatment of an illness or medical condition (e.g., lung cancer, flu or diabetes): they are regarded, following [13], as a major tool in improving the quality of medical care. More concretely, *they describe or define* the "ideal" (most successful) *care plans or therapies* healthcare professionals should follow when treating an "ideal" (i.e., average) patient for a given illness. Guidelines are intented for human consumption. In order however to implement them as clinical workflows or *careflows* within clinical information systems, they have to be translated into machine-readable, executable formal representations of the main control flow features of the described treatment and of its process or plan structure: *computer interpretable guidelines* (CIGs).

The VERICLIG project, funded by a grant from the Free University of Bozen-Bolzano Research Foundation during the period February 2012–January 2014, studied how to automatically extract CIGs from clinical guidelines using natural language processing (NLP) techniques. It was executed, moreover in close collaboration with Claudio Eccher and Elena Cardillo from the eHealth research group of the FBK - Fondazione Bruno Kessler, from Trento, Italy. In this report we summarize the problems addressed during its second and last year (February 2013–January 2014), the NLP techniques considered and their evaluation, plus all the activities, publications and presentations fully or partially supported by VERICLIG.

This report is strutured as follows. Section 2 outlines what we mean by CIGs and the main challenges to tackle when extracting CIGs and fragments thereof via NLP techniques from (English) guidelines. Section 3 provides an overview of the main NLP-inspired information-extraction technique we experimeted with this year: clinical entity and relation recognition. Section 4 then discusses the experiments we made to evaluate this technique, and the conclusions and further work that ensue. Finally, Section 5 lists all the activities and collaborations carried out.

2 Mining Clinical Guidelines

There are several ways to formally characterize processes, but little consensus as to which is the most appropriate for therapies. Thus, we do not intend at this stage to commit ourselves in the VERICLIG project to a particular formalism, but intend rather to focus on the main features such formalisms share, and in particular on their most basic, common constructs. For convenience, we use terminology coming from the Business Process Modeling and Notation (BPMN) standard (see [8]). 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);

[3] `http://www.nice.org.uk/nicemedia/pdf/CG87NICEGuideline.pdf`

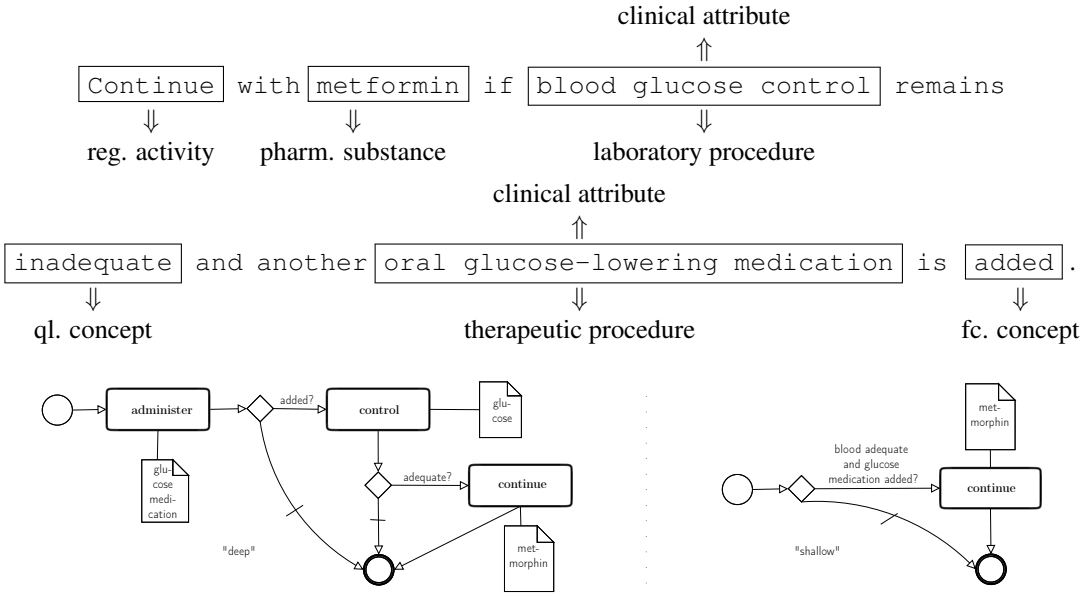


Figure 2: **Top:** MetaMap UMLS (automated) annotations of the NICE diabetes guideline fragment; boxes surround entities, annotations are MetaMap's. **Bottom:** Two candidate CIG fragments (represented in BPMN): to the left, the intended "deep" CIG, to the right a "shallow" CIG. Control flows (diamonds) specify the acceptable orderings of the activities (rounded rectangles); activities consume resources (folded-corner rectangles).

- 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. Such procedure is complex and pitfalls abound. 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. Finally, ambiguity must be resolved. In Figure 2 (top) the reader can see a guideline fragment (recommendation 1.4.1 of the NICE diabetes-2 guideline[1]) with its entities highlighted and their candidate annotations: to correctly extract the "deep" intended CIG fragment (see Figure 2, bottom left) it is necessary to "filter out" the two wrong "clinical attribute" annotations and understand that in this recommendation some nouns refer to activities; we also need to realize that the verb "continue" introduces a third activity, and rely on syntactic structure to properly order the activities.

3 CIG Component Recognition

3.1 Clinical Entity and Relation Recognition

To identify activities and relations in clinical documents, we need to recognize CIG fragments. Let $\vec{t_c} = (c_1, \ldots, c_n)^T$ 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 $\vec{t_r} = (r_1, \ldots, r_n)^T$ of n *relation labels* drawn from a set $\{r_1, \ldots, r_p\}$ of p clinical relations. Let $\vec{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ be a vector of n input *noun phrases* (**NP**s) or *entities* (the n **NP**s of a sentence), or resp., a vector $\vec{\alpha} = ((\alpha_1, \alpha_1), (\alpha_1, \alpha_2), \ldots, (\alpha_n, \alpha_n))^T$ 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,

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

resp., *relation recognition*, see [1], can be formulated as the task of finding the best scoring vector $\vec{t}^*_\beta$:

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

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 $\vec{t}_\beta$ and $\vec{\alpha}$ 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 $\{\text{activity}, \text{resource}, \text{actor}, \text{other}\}$ of entity type labels and the set $\{\text{temporal}, \text{causal}, \text{other}\}$ of relation labels, and consider *supervised* recognizers, viz., recognizers that can be estimated from a training corpus.

3.2 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 [14] we used the Stanford parser (see [7]) 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.

1. By mining parse trees we extracted from **NP**s the following syntactic features: depth *nest* of nesting; position *pos* in the phrase; and occurrence *sub* in a subordinated phrase.

2. The lexical semantic features were extracted by computing several measures of label overlap and frequency. We considered:

   (a) the (raw) frequency of the **NP** entity type c in the corpus;

   (b) 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*; the *relative frequency* φ_{lf} of the **NP** entity type c w.r.t. *labs*; and *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.

4 Experiments and Discussion

In this document we provide a detailed description of our preliminary automated entity recognition experiments, and in particular, the activity recognition experiments. The goal of the experiments was to understand which set of (independent) features semantic (*freq*, *lf*, *hd*, *ls*) or syntactic (*nest*, *pos*, *sub*), see Table 1, has a greater impact for this particular task. We also tried to understand how such features interplay with sentence context and with different types of classifiers (specifically, classifiers tuned to categorical data, such as decision trees, as opposed to classifiers tuned to real-valued features).

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	organization	pharmacological	qualitative
procedure		substance	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

4.1 Gold Corpus

Since no UMLS annotated guideline corpora are available for research purposes we ran our experiments over the SemRep corpus (see [6]), 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. We considered two evaluation strategies:

1. A custom split of SemRep into a training corpus (2/5) and a test or evaluation corpus (3/5).
2. A 10-fold cross-validation, in which the corpus is split 10 times into 10 random subsets (1/10 used for training, 9/10 for evaluation) and the results are then averaged out.

4.2 Experiments

Classifiers. In our experiments the main goal was to evaluate activity recognition features rather than classifier design and evaluation. We thus relied on standard classification models from the known Weka[2] data mining framework. We trained and evaluated the following classifiers:

(i) logistic classifier (Logit), (ii) support vector machine (SVM),
(iii) naive Bayes classifier (Bayes), (iv) neural network (Neural), and
(v) decision tree (Tree).

Context Windows. In parallel to this, we studied the impact of context over activity recognition, and its interplay with our features. To this end we considered a baseline scenario, in which context is restricted to **NP**s, and a scenario in which we take into consideration all the annotated **NP**s of a SemRep sentence. This distinction is important since SemRep is a small and sparsely annotated corpus, for which enhanced feature spaces may not prove informative. These two scenarios were modeled as follows.

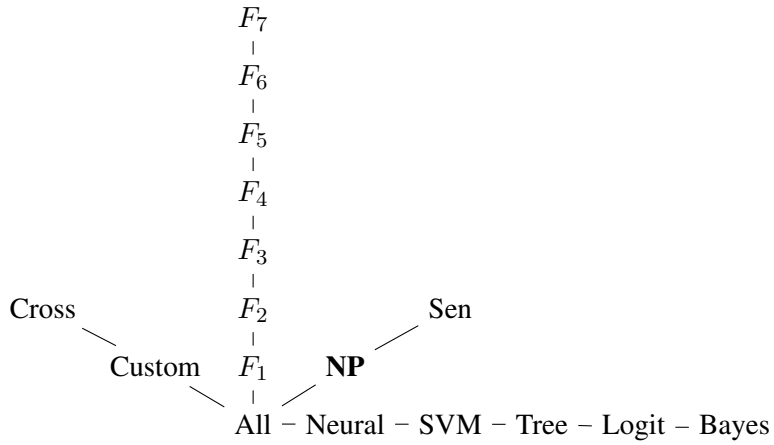
[2] `www.cs.waikato.ac.nz/~ml/weka/`

- A set of **NP** observations: for each **NP** α in SemRep, we extracted feature vector $(f_1^\alpha, \ldots, f_7^\alpha, c^\alpha)^T$.
- A set of sentence observations: for each vector $(\alpha_1, \ldots, \alpha_k)^T$ of annotated **NP**s in a SemRep sentence, we extracted feature vectors $(f_1^{\alpha_1},\ldots,f_7^{\alpha_1},c^{\alpha_1},\ldots,f_1^{\alpha_k},\ldots,f_7^{\alpha_k},c^{\alpha_k})^T$.

Dimensions. To evaluate the impact of the features $\{F_1, \ldots, F_7\}$, we proceded as follows:

1. we removed feature F_i from predictors $\{F_1, \ldots, F_7\}$;
2. we took into account sentence context (Sen scenario) or not (En scenario);
3. we either split the corpus into a training (2/5) and an evaluation (3/5) corpus or considered the whole corpus and
   - if not splitted: evaluated the classifiers via a 10-fold cross-validation over the gold-standard corpus, and measured both individual and average classifier precision (Pr), recall (Re), F1-measure and accuracy (Ac) per each (F, S) feature-scenario pair.
   - if splitted: trained the classifiers over the training corpus, evaluated them over the evaluation corpus, and collected the same performance statistics as before.

In what follows we display the average Pr, Re, F1-measure and Ac results (left), plus the results per classifier (right), and the results by scenario (on top, 10-fold cross-validation and below, the custom split). We also show tables with Pr, Re, F1-measure and Ac results by classifier, feature, context scenario and corpus split. The following diagram summarizes the different dimensions of the experiments:



4.3 Discussion and Conclusions

4.3.1 Discussion

The **NP** scenario shows a drop in average precision, recall, F1-measure and accuracy when *hd* and *freq* are disregarded, and a minor drop when *ls* is disregarded. The removal of syntactic features on the other hand has a smaller effect.

When considering sentence context, we can observe a greater impact for *sub*, and a minor drop when *ls* is disregarded. But sentence context gives rise also to a clear decrease in average classifier performance. Thus *sub*, while significant, is less useful than the semantic features.

Both the 10-cross validation and the custom split settings gave rise to similar trends, as the reader can observe by comparing the plots.

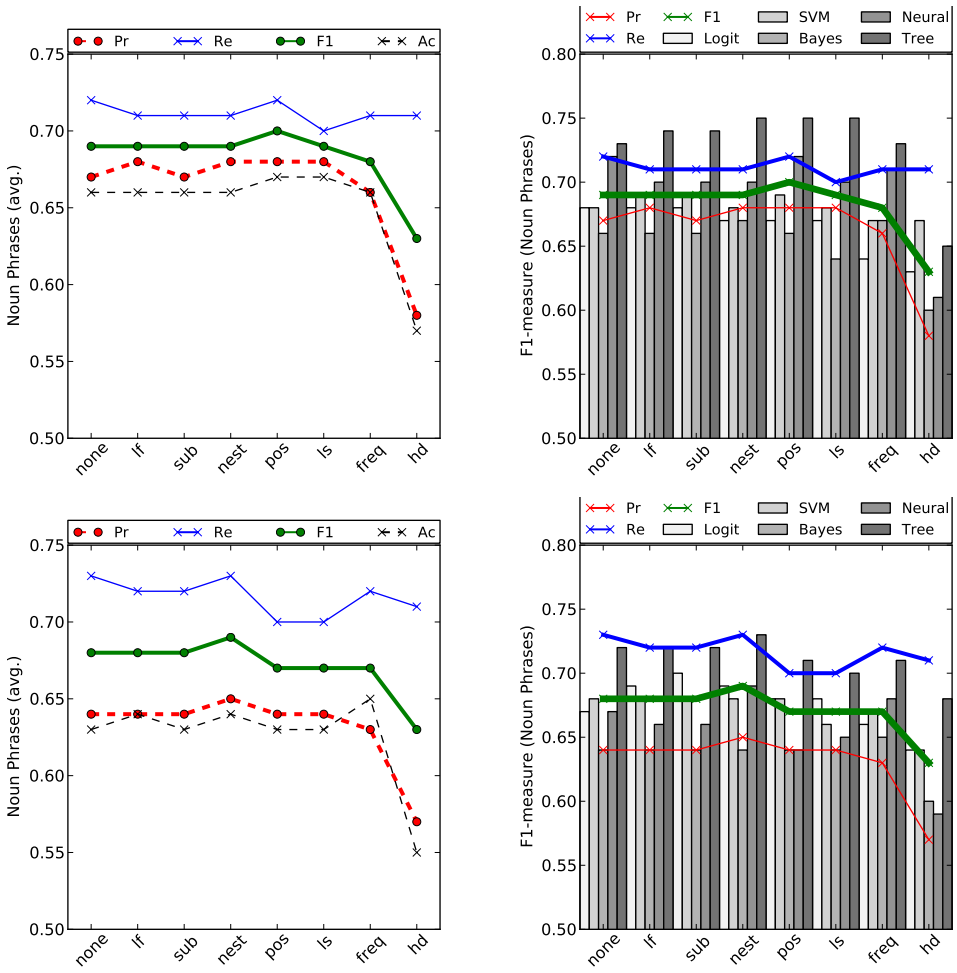


Table 2: **NP** scenario. **Top left:** Average classifier performance, 10-fold cross-validation. **Top right:** Performance per classifier, 10-fold cross validation. **Bottom left:** Average classifier performance, custom split. **Bottom right:** Performance per classifier, custom split.

This last observation is substantiated by corpus evidence. One way to see how, is to focus on the distribution of syntax relatively to corpus domain. Syntactic structures can be approximated by function words[3] (e.g., subordinators such as "if" or "then", coordinators such as "or"). We compared to SemRep:

1. a subset of the Brown corpus ([3]),
2. a corpus of business process specifications ([4]),
3. a subset of the NICE diabetes-2 guideline ([11]),
4. a subset of the NICE eating disorders guideline ([10]), and
5. a subset of the NICE schizophrenia guideline ([12]).

We run the following statistical tests (see [5]) at $p = 0.01$ significance: (i) a t-test (null hypothesis: cross-corpora function word mean relative frequency is 0.20); (ii) a χ^2-test of independence (null hypothesis:

[3]For the POS tagging we relied on a Natural Language Toolkit (NLTK) 3-gram tagger by [2], trained over the (POS annotated) Brown corpus.

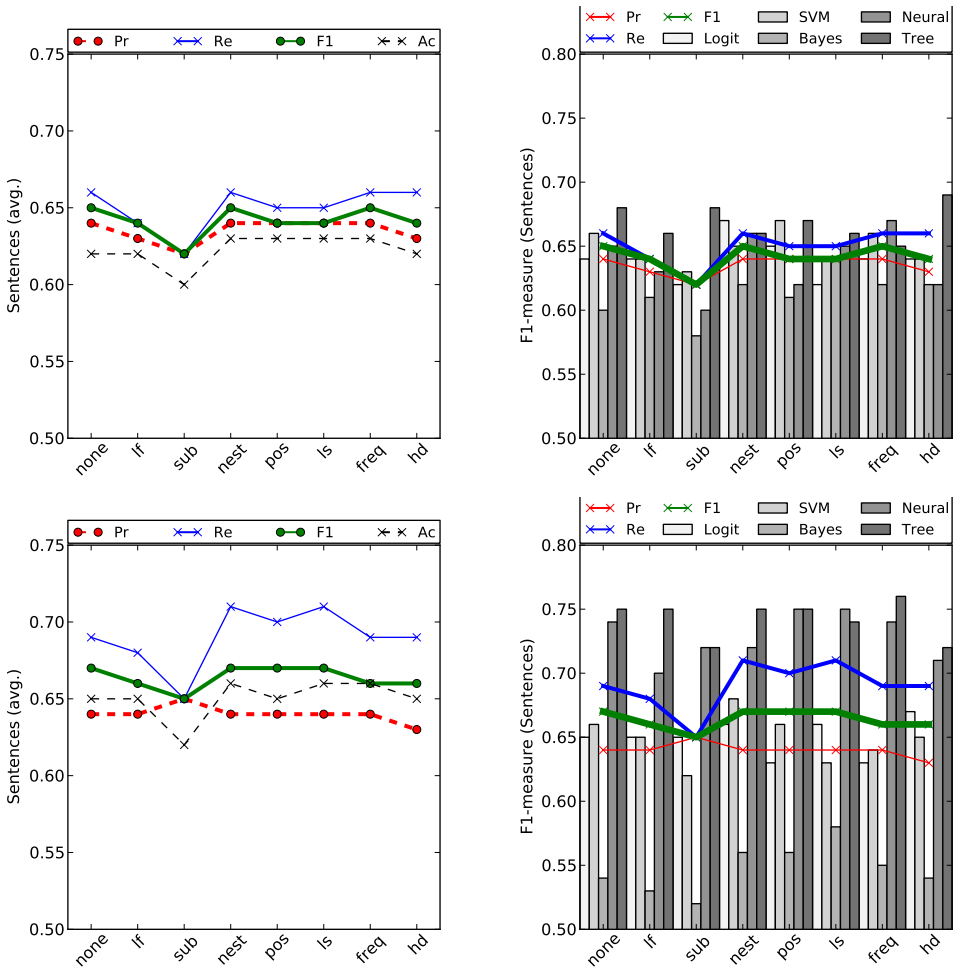


Table 3: Sentence scenario. **Top left:** Average classifier performance, 10-fold cross-validation. **Top right:** Performance per classifier, 10-fold cross validation. **Bottom left:** Average classifier performance, custom split. **Bottom right:** Performance per classifier, custom split.

function word distribution is correlated to corpus domain). The test results (see Table 8) show that syntax is uniform across domains, and thus has a more limited impact relatively to semantics.

Syntax, however, can be leveraged to optimize prediction results when exploited by classifiers sensitive to categorical data. The classifier that performed better overall was the decision tree, which seems to exploit better the more limited impact of *sub*, *pos*, and *nest*.

4.3.2 Conclusions and Further Work

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.

none	Pr	Re	F1	Ac
Logit	0.66	0.69	0.68	0.66
SVM	0.64	0.73	0.68	0.66
Bayes	0.65	0.66	0.66	0.59
Neural	0.66	0.79	0.72	0.68
Tree	0.74	0.73	0.73	0.73

lf	Pr	Re	F1	Ac
Logit	0.66	0.69	0.68	0.66
SVM	0.65	0.73	0.69	0.66
Bayes	0.65	0.67	0.66	0.59
Neural	0.68	0.72	0.70	0.68
Tree	0.74	0.73	0.74	0.72

sub	Pr	Re	F1	Ac
Logit	0.66	0.70	0.68	0.66
SVM	0.65	0.74	0.69	0.67
Bayes	0.65	0.67	0.66	0.59
Neural	0.67	0.73	0.70	0.68
Tree	0.74	0.73	0.74	0.72

nest	Pr	Re	F1	Ac
Logit	0.66	0.68	0.67	0.65
SVM	0.64	0.72	0.68	0.66
Bayes	0.66	0.67	0.67	0.59
Neural	0.67	0.74	0.70	0.67
Tree	0.76	0.74	0.75	0.74

pos	Pr	Re	F1	Ac
Logit	0.65	0.69	0.67	0.66
SVM	0.64	0.73	0.69	0.66
Bayes	0.66	0.67	0.66	0.59
Neural	0.70	0.75	0.72	0.69
Tree	0.74	0.76	0.75	0.73

ls	Pr	Re	F1	Ac
Logit	0.67	0.68	0.67	0.66
SVM	0.67	0.69	0.68	0.66
Bayes	0.63	0.65	0.64	0.61
Neural	0.68	0.72	0.70	0.67
Tree	0.75	0.76	0.75	0.73

freq	Pr	Re	F1	Ac
Logit	0.63	0.66	0.64	0.64
SVM	0.62	0.72	0.67	0.64
Bayes	0.64	0.71	0.67	0.64
Neural	0.66	0.76	0.71	0.68
Tree	0.73	0.72	0.73	0.71

hd	Pr	Re	F1	Ac
Logit	0.57	0.71	0.63	0.57
SVM	0.56	0.82	0.67	0.58
Bayes	0.57	0.64	0.60	0.53
Neural	0.58	0.65	0.61	0.57
Tree	0.60	0.72	0.65	0.60

avg	Pr	Re	F1	Ac
none	0.67	0.72	0.69	0.66
lf	0.68	0.71	0.69	0.66
sub	0.67	0.71	0.69	0.66
nest	0.68	0.71	0.69	0.66
pos	0.68	0.72	0.70	0.67
ls	0.68	0.70	0.69	0.67
freq	0.66	0.71	0.68	0.66
hd	0.58	0.71	0.63	0.57

Table 4: Precision, recall, F1-measure and accuracy per classifier and feature. **NP** scenario, 10-fold cross-validation.

none	Pr	Re	F1	Ac
Logit	0.63	0.72	0.67	0.64
SVM	0.60	0.78	0.68	0.63
Bayes	0.65	0.64	0.64	0.55
Neural	0.60	0.77	0.67	0.64
Tree	0.70	0.76	0.72	0.69

lf	Pr	Re	F1	Ac
Logit	0.64	0.74	0.69	0.65
SVM	0.60	0.78	0.68	0.63
Bayes	0.64	0.63	0.64	0.57
Neural	0.63	0.70	0.66	0.64
Tree	0.70	0.76	0.72	0.69

sub	Pr	Re	F1	Ac
Logit	0.65	0.76	0.70	0.66
SVM	0.60	0.77	0.68	0.63
Bayes	0.64	0.63	0.64	0.54
Neural	0.62	0.70	0.66	0.63
Tree	0.70	0.76	0.72	0.69

nest	Pr	Re	F1	Ac
Logit	0.64	0.74	0.69	0.65
SVM	0.61	0.78	0.68	0.64
Bayes	0.64	0.64	0.64	0.55
Neural	0.65	0.74	0.69	0.66
Tree	0.70	0.77	0.73	0.69

pos	Pr	Re	F1	Ac
Logit	0.63	0.72	0.68	0.64
SVM	0.60	0.78	0.68	0.63
Bayes	0.65	0.63	0.64	0.55
Neural	0.65	0.63	0.64	0.64
Tree	0.69	0.74	0.71	0.68

ls	Pr	Re	F1	Ac
Logit	0.65	0.72	0.68	0.65
SVM	0.57	0.78	0.66	0.62
Bayes	0.64	0.64	0.64	0.56
Neural	0.63	0.67	0.65	0.63
Tree	0.71	0.69	0.70	0.68

freq	Pr	Re	F1	Ac
Logit	0.62	0.71	0.66	0.64
SVM	0.60	0.75	0.67	0.63
Bayes	0.64	0.65	0.65	0.62
Neural	0.61	0.77	0.68	0.66
Tree	0.70	0.73	0.71	0.69

hd	Pr	Re	F1	Ac
Logit	0.57	0.72	0.64	0.57
SVM	0.55	0.75	0.64	0.56
Bayes	0.57	0.62	0.60	0.48
Neural	0.57	0.60	0.59	0.54
Tree	0.57	0.85	0.68	0.61

avg	Pr	Re	F1	Ac
none	0.64	0.73	0.68	0.63
lf	0.64	0.72	0.68	0.64
sub	0.64	0.72	0.68	0.63
nest	0.65	0.73	0.69	0.64
pos	0.64	0.70	0.67	0.63
ls	0.64	0.70	0.67	0.63
freq	0.63	0.72	0.67	0.65
hd	0.57	0.71	0.63	0.55

Table 5: Precision, recall, F1-measure and accuracy per classifier and feature. **NP** scenario, custom split.

none	Pr	Re	F1	Ac
Logit	0.66	0.62	0.64	0.61
SVM	0.62	0.71	0.66	0.65
Bayes	0.61	0.59	0.60	0.58
Neural	0.63	0.67	0.65	0.62
Tree	0.66	0.70	0.68	0.66

lf	Pr	Re	F1	Ac
Logit	0.63	0.64	0.64	0.62
SVM	0.60	0.68	0.64	0.63
Bayes	0.62	0.61	0.61	0.59
Neural	0.64	0.63	0.63	0.63
Tree	0.66	0.65	0.66	0.65

sub	Pr	Re	F1	Ac
Logit	0.61	0.63	0.62	0.60
SVM	0.61	0.65	0.63	0.60
Bayes	0.60	0.55	0.58	0.55
Neural	0.60	0.61	0.60	0.58
Tree	0.68	0.68	0.68	0.65

nest	Pr	Re	F1	Ac
Logit	0.67	0.66	0.67	0.65
SVM	0.60	0.69	0.65	0.63
Bayes	0.63	0.61	0.62	0.59
Neural	0.65	0.68	0.66	0.63
Tree	0.66	0.65	0.66	0.65

pos	Pr	Re	F1	Ac
Logit	0.67	0.64	0.65	0.63
SVM	0.63	0.71	0.67	0.66
Bayes	0.62	0.61	0.61	0.59
Neural	0.62	0.63	0.62	0.60
Tree	0.66	0.68	0.67	0.66

ls	Pr	Re	F1	Ac
Logit	0.63	0.62	0.62	0.62
SVM	0.61	0.69	0.64	0.64
Bayes	0.65	0.63	0.64	0.62
Neural	0.67	0.63	0.65	0.62
Tree	0.65	0.66	0.66	0.65

freq	Pr	Re	F1	Ac
Logit	0.65	0.64	0.64	0.63
SVM	0.62	0.69	0.66	0.64
Bayes	0.62	0.62	0.62	0.61
Neural	0.63	0.71	0.67	0.63
Tree	0.66	0.64	0.65	0.65

hd	Pr	Re	F1	Ac
Logit	0.64	0.64	0.64	0.60
SVM	0.61	0.69	0.64	0.64
Bayes	0.62	0.62	0.62	0.60
Neural	0.60	0.64	0.62	0.58
Tree	0.68	0.71	0.69	0.66

avg	Pr	Re	F1	Ac
none	0.64	0.66	0.65	0.62
lf	0.63	0.64	0.64	0.62
sub	0.62	0.62	0.62	0.60
nest	0.64	0.66	0.65	0.63
pos	0.64	0.65	0.64	0.63
ls	0.64	0.65	0.64	0.63
freq	0.64	0.66	0.65	0.63
hd	0.63	0.66	0.64	0.62

Table 6: Precision, recall, F1-measure and accuracy per classifier and feature. Sentence scenario, 10-fold cross-validation.

none	Pr	Re	F1	Ac
Logit	0.62	0.69	0.65	0.63
SVM	0.60	0.73	0.66	0.64
Bayes	0.56	0.51	0.54	0.54
Neural	0.73	0.74	0.74	0.73
Tree	0.71	0.79	0.75	0.73

lf	Pr	Re	F1	Ac
Logit	0.62	0.68	0.65	0.64
SVM	0.59	0.73	0.65	0.63
Bayes	0.55	0.51	0.53	0.53
Neural	0.72	0.69	0.70	0.72
Tree	0.71	0.79	0.75	0.73

sub	Pr	Re	F1	Ac
Logit	0.65	0.65	0.65	0.61
SVM	0.58	0.66	0.62	0.57
Bayes	0.58	0.47	0.52	0.53
Neural	0.76	0.69	0.72	0.70
Tree	0.66	0.79	0.72	0.70

nest	Pr	Re	F1	Ac
Logit	0.61	0.71	0.66	0.64
SVM	0.60	0.78	0.68	0.65
Bayes	0.59	0.53	0.56	0.57
Neural	0.70	0.73	0.72	0.72
Tree	0.71	0.79	0.75	0.73

pos	Pr	Re	F1	Ac
Logit	0.61	0.66	0.63	0.63
SVM	0.60	0.73	0.66	0.63
Bayes	0.57	0.54	0.56	0.55
Neural	0.73	0.77	0.75	0.73
Tree	0.70	0.80	0.75	0.73

ls	Pr	Re	F1	Ac
Logit	0.65	0.67	0.66	0.65
SVM	0.57	0.71	0.63	0.61
Bayes	0.59	0.58	0.58	0.57
Neural	0.71	0.79	0.75	0.73
Tree	0.69	0.81	0.74	0.72

freq	Pr	Re	F1	Ac
Logit	0.62	0.65	0.63	0.64
SVM	0.58	0.71	0.64	0.62
Bayes	0.57	0.53	0.55	0.55
Neural	0.71	0.76	0.74	0.73
Tree	0.72	0.81	0.76	0.74

hd	Pr	Re	F1	Ac
Logit	0.64	0.72	0.67	0.64
SVM	0.56	0.75	0.65	0.61
Bayes	0.51	0.57	0.54	0.54
Neural	0.71	0.71	0.71	0.72
Tree	0.74	0.70	0.72	0.73

avg	Pr	Re	F1	Ac
none	0.64	0.69	0.67	0.65
lf	0.64	0.68	0.66	0.65
sub	0.65	0.65	0.65	0.62
nest	0.64	0.71	0.67	0.66
pos	0.64	0.70	0.67	0.65
ls	0.64	0.71	0.67	0.66
freq	0.64	0.69	0.66	0.66
hd	0.63	0.69	0.66	0.65

Table 7: Precision, recall, F1-measure and accuracy per classifier and feature. Sentence scenario, custom split.

corpus	size (words)	domain	rel. freq.
Brown	1,391,708	news	0.16
Friederich	3,824	processes	0.17
SemRep	13,948	clinical	0.18
diabetes2	7,109	clinical	0.16
eating dis.	5,078	clinical	0.17
schizophrenia	5,367	clinical	0.18

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

Table 8: **Top:** Function word relative frequency across corpora and domains. **Bottom:** Statistical tests (χ^2-test of independence and t-test).

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 [1]. 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 [9].

5 Publications, Presentations and Collaborations

- Tutorials, Posters, Presentations:

  1. CSLT 2013, Potsdam, Germany, March 2013. (Presentation)
  2. ESSLLI 2013, Düsseldorf, Germany, August 2013. (Tutorial)
  3. TbiLLC 2013, Gudauri, Georgia, September 2013. (Presentation)
  4. IJCNLP 2013, Nagoya, Japan, October 2013. (Poster)
  5. AI*IA 2013, Torino, Italy, December 2013. (Presentation)

- (Peer-reviewed) Publications:

  1. Camilo Thorne, Elena Cardillo, Claudio Eccher, Marco Montali, Diego Calvanese. "The VERICLIG Project: Extraction of Computer Interpretable Guidelines via Syntactic and Semantic Annotation". 2013. Proceedings of CSLT 2013, p. 54–58, ISBN 978-1-6274839-8-8. (Workshop paper)
  2. Camilo Thorne, Elena Cardillo, Claudio Eccher, Marco Montali, Diego Calvanese. "Automated Activity Recognition in Clinical Documents". 2013. Proceedings of IJCNLP 2013, p. 1129–1133, ISBN 978-4-9907348-0-0. (Conference paper)

3. Camilo Thorne, Elena Cardillo, Claudio Eccher, Marco Montali, Diego Calvanese. "Process Fragment Recognition in Clinical Documents". 2013. Proceedings of AI*IA 2013, p. 227–238, ISBN 978-3-3190352-3-9. (Conference paper)
4. Camilo Thorne, Raffaella Bernardi, Diego Calvanese. "Designing Efficient Controlled Languages for Ontologies". 2014. Computing Meaning, p. 149–176, ISBN 978-9-4007728-4-7. (Book chapter)
5. Camilo Thorne, Jakub Szymanik. "Semantic Complexity of Quantfiers and Quantifier Distribution in Corpora". 2014. Proceedings of TbiLLC2013. Submitted. (Symposium paper)

- Collaborations:

1. E. Cardillo and C. Eccher, FBK, eHealth Group. (Trento)
2. K. Kaiser, TUWien. (Wien)
3. M. Arguello, UManchester. (Manchester)
4. J. Szymanik, UVA. (Amsterdam)
5. R. Bernardi, UniTN. (Trento)

References

[1] Asma Ben Abacha and Pierre Zweigenbaum. Medical entity recognition: A comparison of semantic and statistical methods. In *Proceedings of the BioNLP 2011 Workshop*, 2011.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009.

[3] Nelson Francis and Henry Kucera. A standard corpus of present-day edited american english, for use with digital computers. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, USA, 1964.

[4] Fabian Friederich, Jan Mendling, and Frank Puhlmann. Process model generation from natural language text. In *Proc. of the 23rd Int. Conf. on Advanced Information Systems Engineering (CAiSE 2011)*, 2011.

[5] Stefan Th. Gries. Useful statistics for corpus linguistics. In Aquilino Sánchez and Moisés Almela, editors, *A mosaic of corpus linguistics: selected approaches*, pages 269–291. Peter Lang, 2010.

[6] Halil Kilicoglu, Graciela Rosenblat, Marcelo Fiszman, and Thomas C. Rindfleisch. Constructing a semantic predication gold standard from the biomedical literature. *BMC Bioinformatics*, 12(486), 2011.

[7] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics ACL 2003*, 2003.

[8] Ryan K.L. Ko, Stephen S.G. Lee, and Eng Wah Lee. Business process mangament (BPM) standards: A survey. *Business Process Management J.*, 15(5):744–791, 2009.

[9] Nataliya Mulyar, Maja Pesic, Wil M. P. van der Aalst, and Mor Peleg. Declarative and procedural approaches for modelling clinical guidelines: Adressing fexibility issues. In *Proc. of the 5th Int. Conf. on Business Process Modelling (BPM 2007)*, 2007.

[10] NICE - NHS. *Eating Dissorders*. National Institute for Health and Clinical Excellence (UK), 2004. Available from http://www.nice.org.uk/nicemedia/live/10932/29218/29218.pdf.

[11] NICE - NHS. *Type 2 Diabetes*. National Institute for Health and Clinical Excellence (UK), 2008. Available from http://www.nice.org.uk/nicemedia/pdf/CG87NICEGuideline.pdf.

[12] NICE - NHS. *Schizophrenia*. National Institute for Health and Clinical Excellence (UK), 2010. Available from http://www.nice.org.uk/nicemedia/pdf/CG87NICEGuideline.pdf.

[13] Yuval Shahar, Ohad Young, Erez Shalom, Maya Glaperin, Alon Mayafitt, Robert Moskovitch, and Alon Hessing. A framework for a distributed, hybrid, multiple-ontology clinical-guideline library and automated guideline-support tools. *J. of Biomedical Informatics*, 37(5):325–344, 2004.

[14] Deyu Zhou and Yulan He. Semantic parsing for biomedical event extraction. In *Proc. of the Ninth Int. Conf. on Computational Semantics IWCS 2011*, 2011.