Automated Activity Recognition in Clinical Documents

C. Thorne M. Montali D. Calvanese
Free University of Bozen-Bolzano
Bolzano, Italy
surname@inf.unibz.it

E. Cardillo C. Eccher
Fondazione Bruno Kessler
Trento, Italy
surname@fbk.edu

Abstract

We describe a first experiment on the identification and extraction of computer-interpretable guideline (CIG) components (activities, actors and consumed artifacts) from clinical documents, based on clinical entity recognition techniques. We rely on MetaMap and the UMLS Metathesaurus to provide lexical information, and study the impact of clinical document syntax and semantics on activity recognition.

Keywords. Clinical entity recognition, computer interpretable guideline, UMLS Metathesaurus.

Introduction. Clinical practice guidelines are systematically developed documents specifying the activities, resources and personnel required to cure or treat an specific illness or medical condition (Field and Lohr (1990)). The need to instantiate them into clinical protocols and workflows has given rise to *computer-interpretable guidelines* (CIGs) (De Clercq et al. (2008)), i.e., formal representations of the care process or plan, and to several natural language processing (NLP) techniques aimed at automating the costly manual CIG generation process (Kaiser et al. (2007), Serban et al. (2007)). All NLP approaches leverage on annotated biomedical resources (e.g., the CLEF corpus from Roberts et al. (2007) and Mykowiecka and Marciniak (2011)), or on frameworks such as cTAKES (Savova et al. (2010)). The key lexical-semantic resource in this domain is the US National Library of Medicine's Unified Medical Language System (UMLS) Metathesaurus (Bodenreider (2004)), complemented by its front-end MetaMap (Aronson and Lang (2010)).

In this paper we conduct a first experiment on how to apply entity recognition techniques inspired by Abacha and Zweigenbaum (2011), to recognize CIG components in medical documents. The process dimension of CIGs consists of four pillars: **(1)** *activities* to be executed; **(2)** the *resources* they use or consume; **(3)** the *actors* that execute them; **(4)** *control flows and gates* that temporally constrain activities. We focus in this paper on activities, the main building block of CIGs, and to a lesser extent on resources and actors. All these components are denoted by content words and can be used to build CIG fragments. We rely on MetaMap annotations and evaluate our techniques over an UMLS-annotated clinical corpus.

CIGs and Activities. Activities are entities difficult to identify with current resources: within clinical documents, in fact, not only verbs (**VB**s) but also proper nouns (**PN**s), common nouns (**NN**s) and, more in general, noun phrases (**NP**s)[1] can refer to them. Figure 1 shows an example from the type-2 diabetes guideline of the National Institute for Health and Clinical Excellence (NICE) (NICE - NHS (2009)) expressing a conditional CIG/process fragment, annotated automatically with MetaMap. To correctly extract the "deep" intended representations it is necessary to recognize that the two entities "blood glucose control" and "oral glucose-lowering medication" are activity tokens. MetaMap annotations provide a clue, but we still need to "filter out" the "clinical attribute" UMLS annotation. We want to understand how this information can be used for this task within an entity recognition framework.

Clinical Entity Recognition. Let $\vec{c}$ denote a vector of clinical *entity type labels*, and $\vec{\alpha}$ a vector of input *noun phrases* (**NP**s) or *entities*. The goal of *clinical entity recognition*, see Abacha and Zweigenbaum (2011), can be formulated as the task of finding the best scoring vector of clinical

[1] In this paper we refer to the Penn Treebank part-of-speech (POS) notation as described by Marcus et al. (1993).

International Joint Conference on Natural Language Processing, pages 1129–1133,
Nagoya, Japan, 14-18 October 2013.

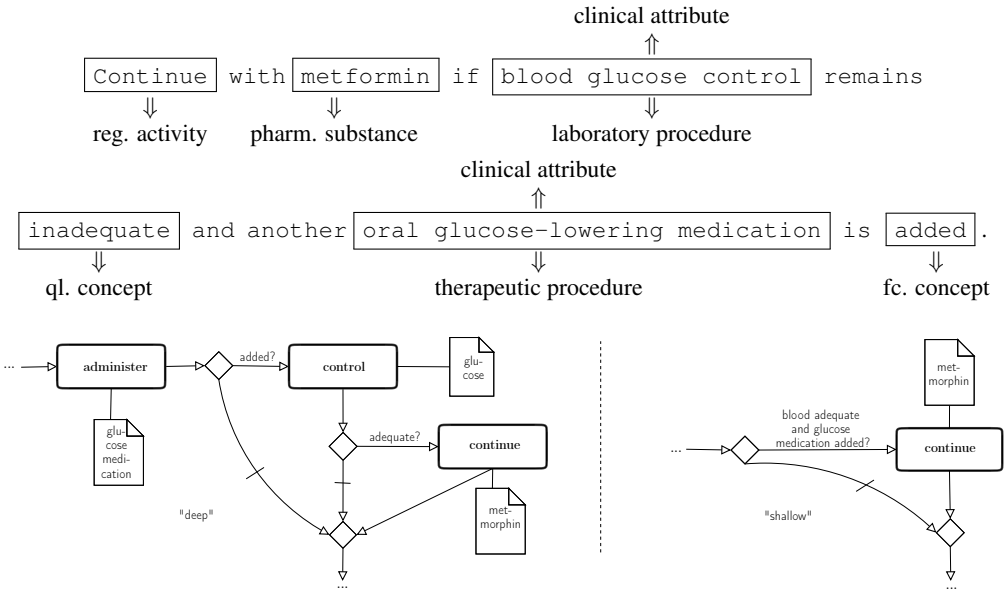


Figure 1: **Top:** MetaMap UMLS (automated) annotations of the NICE diabetes guideline fragment; boxes surround entities, annotations are MetaMap's. **Bottom:** Two candidate CIG fragments (represented in Business Process Modeling Notation (BPMN), see Ko et al. (2009)): to the left, the intended "deep" 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).

entity type labels: $\vec{c}^* = \arg\max\{\vec{c} \mid \mu(\rho(\vec{\alpha}, \vec{c}))\}$, where $\mu(\cdot)$ denotes a *recognizer* built using a classification model (e.g., a logistic regression algorithm), and $\rho(\cdot, \cdot)$ is a feature extraction function. In the following paragraphs we study this task w.r.t. the set $\{\text{activity}, \text{resource}, \text{actor}, \text{other}\}$ of entity types.

The SemRep corpus. Since no UMLS annotated clinical guideline corpora are available for research purposes, we ran our experiments over the SemRep corpus by Kilicoglu et al. (2011), a small annotated clinical corpus whose domain largely overlaps with that of guidelines. 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.

Features. The focus of our experiments is to understand the predictive power of syntax and semantics for CIG entity recognition, and in particular for activity recognition. Intuitively, both syntax and semantics can contribute to the prediction of clinical entity types, but it is not a priori clear which one contributes more. Similarly to Zhou and He (2011) we used the Stanford parser (see Klein and Manning (2003)) to extract syntactic features, and MetaMap to extract semantic features. We harvested clinical types by mapping UMLS concept types returned by MetaMap to their subsuming clinical types. In the top of Table 1 we show a sample of UMLS concept types subsumed by "activity", "resource", "actor" and "other", whereas in its bottom we summarize the extracted features, described in detail below.

By mining the **NP**s sentence parse trees, we extracted the following syntactic features: depth of nesting (*nest*); position in the phrase (*pos*); occurrence in a subordinated phrase (*sub*). The intuition behind these features is that certain types may correlate strongly with syntax (e.g., one would expect "resource" to annotate an object **NP**).

The semantic features were extracted by computing several measures of label overlap and frequency. The rationale of these features is that, while MetaMap outputs many possible clinical meanings of the constituent **NN**s of an **NP** entity, giving rise to multiple "activity", "resource", "actor" and "other" annotations per **NN** and **NP**, it tends to output meanings that are semantically related (within the UMLS Metathesaurus hierarchy) to the **NP**'s intended type.

We measured the raw frequency *freq* of the **NP** entity type c in the SemRep corpus, the degree of annotation overlap *hd* between the bag of possi-

activity	actor	resource	other
laboratory procedure	professional society	manufactured object	qualitative concept

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 $\in \mathbb{N}$
lf	rel. freq. of label in **NP**	real $\in [0, 1]$
hd	head/**NP** overlap	real $\in [0, 1]$
ls	label/**NP** overlap	real $\in [0, 1]$
class	**NP** entity type	act., actor, res., other

Table 1: **Top:** CIG entity labels and sample UMLS concept types they subsume. **Bottom: NP** features considered; the class label is the *dependent* feature we want to predict.

bly repeated labels *labs* collected using MetaMap from all the **NN**s in an **NP**, and the bag of possibly repeated labels of its head noun *labsh*. In addition, we computed the relative frequency *lf* of the **NP** entity type c w.r.t. *labs*:

$$hd = \frac{||labs \Cap labsh||}{||labs|| + ||labsh||} \quad lf = \frac{||labs \Cap \{c\}||}{||labs||} \qquad (1)$$

where $||\cdot||$ and $\Cap$ denote resp. bag cardinality and intersection. The intuition behind these two features is that the intended type will tend to prevail within the annotations of an **NP**, and in particular among its head **NN** and its modifiers. Finally, we took into account the taxonomical structure of the UMLS Metathesaurus and defined the following label/**NP** overlap *ls*:

$$ls = \frac{||labs \Cap sub(c)||}{||labs|| + ||sub(c)||} \qquad (2)$$

where $sub(c)$ is the bag of all the UMLS concept types that are subsumed by the entity type label c. The *ls* feature measures how similar are the MetaMap **NP** annotations to the UMLS hierarchy subsumed by c. In all cases, a simple Laplace smoothing was applied.

Evaluation Framework. 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). To measure the significance of each single feature, we removed each time a feature F_i from the space $\{F_1, \ldots, F_7\}$ of syntactic and semantic *independent* features from Table 1 and retrained and reevaluated the classifiers w.r.t. the feature space $\{F_1, \ldots, F_{i-1}, F_{i+1}, \ldots, F_7\}$.

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. **(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 annotated **NP**s in a SemRep sentence, we extracted feature vectors $(f_1^{\alpha_1}, \ldots, f_7^{\alpha_1}, c^{\alpha_1}, \ldots, f_1^{\alpha_k}, \ldots, f_7^{\alpha_k}, c^{\alpha_k})^T$.

For each combination of classifier feature and scenario, we performed a 10-fold cross-validation to measure precision (Pr), recall (Re), F1-measure, and the overall accuracy (Ac) of the classifiers for the activity recognition task[3].

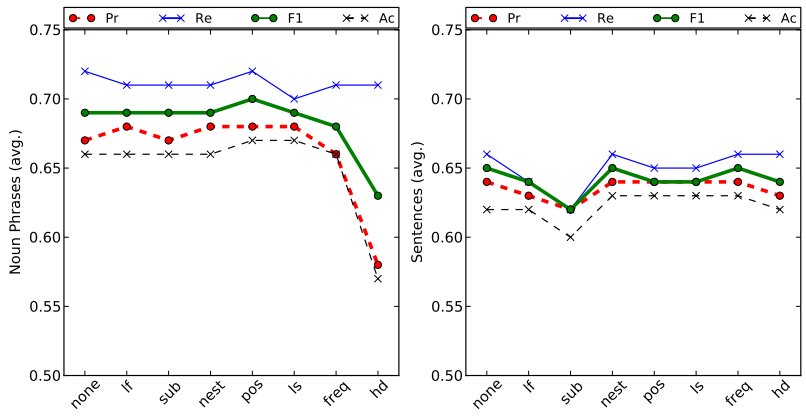
Results and Discussion. The baseline scenario (see Figure 2, left) shows a drop in average precision, recall, F-measure and accuracy when *hd* and *freq* are disregarded, and a minor drop when *ls* is disregarded. The removal of syntactic features on the other hand has a smaller effect. Considering sentence context (see Figure 2, center), 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.

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[4] (e.g., subordinators (**IN**s) such as "if"

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

[3] For reasons of space, we present here a summary of the results obtained; for a more detailed description, please refer to www.inf.unibz.it/~cathorne/vericlig/ijcnlp2013-exp.pdf

[4] For the POS tagging we relied on a Natural Language Toolkit (NLTK) 3-gram tagger by Bird et al. (2009), trained over the (POS annotated) Brown corpus.



		Pr	Re	F1	Ac
Logit	(**NP**)	0.66	0.69	0.68	0.66
	(sen.)	0.66	0.62	0.64	0.61
SVM	(**NP**)	0.64	0.73	0.68	0.66
	(sen.)	0.62	0.71	0.66	0.65
Bayes	(**NP**)	0.65	0.66	0.66	0.59
	(sen.)	0.61	0.59	0.60	0.58
Neural	(**NP**)	0.66	0.79	0.72	0.68
	(sen.)	0.63	0.67	0.65	0.62
Tree	(**NP**)	0.74	0.73	0.73	0.73
	(sen.)	0.66	0.70	0.68	0.66

Figure 2: **Left, Center:** Results of 10-fold cross-validation by scenario. On the y-axis, activity recognition precision, recall, F1-measure and classifier accuracy (classifier averages). On the x-axis, the feature(s) removed. The tag "none" means that no feature was removed. **Right:** Results for the original (complete) feature space, by classifier and label context (noun phrase **NP** or sentence *sen.*).

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

or "then", coordinators (**CC**s) such as "or").

We compared to SemRep: *(i)* a subset of the Brown corpus (Francis and Kucera (1964)), *(ii)* a corpus of business process specifications (Friederich et al. (2011)), *(iii)* a subset of the NICE diabetes-2 guideline (NICE - NHS (2009)), *(iv)* a subset of the NICE eating disorders guideline (NICE - NHS (2004)), and *(v)* a subset of the NICE schizophrenia guideline (NICE - NHS (2010)). We run the following statistical tests (see Gries (2010)) at $p = 0.01$ significance: **(1)** a t-test (null hypothesis: cross-corpora function word mean relative frequency is 0.20); **(2)** a χ^2-test of independence (null hypothesis: function word distribution is correlated to corpus domain). The test results (see Table 2) 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 (see Figure 2, right), which seems to exploit better the more limited impact of *sub*, *pos*, and *nest*.

Conclusions and Further Work. We have conducted preliminary experiments on automatic clinical activity recognition using MetaMap and entity recognition techniques. We experimented our techniques on the SemRep gold standard UMLS-annotated corpus. Our experiments suggest that the semantic environment of an entity is more useful for this task. Corpus analysis on SemRep and other corpora seems to confirm this observation. In the future, we plan to consider more powerful classification models for NLP, such as conditional random fields (CRFs), able to exploit possible dependencies among features. We plan to focus on document semantics, by considering more complex semantic features (based on, e.g., thesaurus-based similarity metrics). Finally, to better cope with data sparseness we intend to consider a bigger corpus by integrating SemRep with, e.g., the i2b2 clinical corpus as suggested by Abacha and Zweigenbaum (2011).

Acknowledgments. The present work has been done within the context of the VERICLIG project[5], supported by a grant from the Free University of Bozen-Bolzano Foundation.

[5] www.unibz.it/~cathorne/vericlig

References

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

Alan R. Aronson and François-Michel Lang. 2010. And overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

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

Olivier Bodenreider. 2004. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32:D267D270.

Paul De Clercq, Katharina Kaiser, and Arie Hasman. 2008. Computer interpretable medical guidelines. In A. Ten Teije et al., editor, *Computer-based Medical Guidelines and Protocols: A Primer and Current Trends*, chapter 2, pages 22–43. IOS Press.

Marilyn J. Field and Kathleen N. Lohr, editors. 1990. *Clinical Practice Guidelines. Directions for a New Program*. National Academy Press.

Nelson Francis and Henry Kucera. 1964. 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.

Fabian Friederich, Jan Mendling, and Frank Puhlmann. 2011. Process model generation from natural language text. In *Proceedings of the 23rd International Conference on Advanced Information Systems Engineering (CAiSE 2011)*.

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

Katharina Kaiser, Cem Akaya, and Silvia Miksch. 2007. 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.

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

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

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

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

Agnieszka Mykowiecka and Malgorzata Marciniak. 2011. Some remarks on automatic semantic annotation of a medical corpus. In *Proceedings of the 3rd International Workshop on Health Document Text Mining and Information Systems*.

NICE - NHS. 2004. Eating dissorders. Available from `http://www.nice.org.uk/nicemedia/live/10932/29218/29218.pdf`.

NICE - NHS. 2009. Type 2 diabetes. Available from `http://www.nice.org.uk/nicemedia/pdf/CG87NICEGuideline.pdf`.

NICE - NHS. 2010. Schizophrenia. Available from `http://www.nice.org.uk/nicemedia/pdf/CG87NICEGuideline.pdf`.

Angus Roberts, Robert Gaizaskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, and Bill Wheeldin. 2007. The CLEF corpus: Semantic annotation of a clinical text. In *Proceedings of the AMIA 2007 Annual Symposium*.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Radu Serban, Anette ten Teije, Frank van Harmelen, Mar Marcos, and Cristina Polo-Conde. 2007. Extraction and use of linguistics patterns for modelling medical guidelines. *Artificial Intelligence in Medicine*, 39(2):137–149.

Deyu Zhou and Yulan He. 2011. Semantic parsing for biomedical event extraction. In *Proceedings of the 9th International Conference on Computational Semantics IWCS 2011*.