

VERICLIG: Extraction and Verification of Clinical Guidelines

Project Report for 2012

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

KRDB Research Centre for Knowledge and Data

Piazza Domenicani 3, Bolzano, Italy

{cthorne,montali,calvanese}@inf.unibz.it

February 1, 2013

Contents

Introduction	1
1 Clinical Guidelines and Processes	2
1.1 Processes	2
1.2 Process-evoking Words (PEWs)	3
1.3 Negative Polarity-evoking Words (NEGs)	3
2 Extraction of CIGs	3
2.1 Limitations of Current Biomedical Resources	4
2.2 Business Process Extraction	5
2.3 Clinical Word Sense Disambiguation	6
2.4 Methodology	6
3 Guideline Analysis	6
3.1 Pattern-based Analysis	7
3.2 Hypothesis Testing	8
3.3 Results and Discussion.	9
4 Results for 2012 and Work Plan for 2013	10
4.1 Results for 2012	10
4.2 Plan for 2013	10
References	10

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 [12], as a major tool in improving the quality of medical care. More concretely, *they describe*

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 guideline³. Each line describes atomic treatments that combine together into a complex therapy.

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 illness¹.

Guidelines need to be modified or instantiated relatively to available resources by health institutions, patients or doctors into protocols, and implemented thereafter into clinical workflows or *careflows* within clinical information systems. An important intermediate step for the synthesis of protocols and careflows from guidelines are *computer interpretable guidelines* (CIGs), viz., formal representations of the main control flow features of the described treatment and of its process or plan structure. CIGs can be exploited in a plethora of ways by clinical decision support systems to provide execution support and recommendations to the involved practitioners, guide the refinement into executable clinical protocols and careflows, and check for conformance and compliance.

The VERICLIG project², intends to address this problem by adopting a natural language processing (NLP) and, in particular, a computational semantics approach that aims at extracting CIGs from textual clinical guidelines. Our objective is to extract the main control-flow structures emerging from the textual description of guidelines in order to explore, in a second step, the possibility to express them using well-known representation languages. One such language is the business processing modeling notation (BPMN) standard (see [6]).

1 Clinical Guidelines and Processes

Clinical guidelines such as, for instance, guidelines related to chronic diseases such as diabetes, allergies or lactose intolerance, are minimally structured documents. They possess however some crucial features: **(1)** they describe a *process*, generically intended as a set of coordinated activities, structured over time, to jointly reach a certain goal, and **(2)** the structure of the process they describe is significantly reflected by English *syntax* and *vocabulary*.

1.1 Processes

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 the terminology coming from the BPMN standard. In BPMN a process is a complex object constituted by the following basic components:

¹The definition of the problem studied in VERICLIG and the results thus far obtained would not have been possible without the collaboration of Claudio Eccher and Elena Cardillo from the eHealth research group of the FBK - Fondazione Bruno Kessler, at Povo, Italy. The VERICLIG project is supported by a grant from the Free University of Bozen-Bolzano Foundation.

²<http://www.inf.unibz.it/~cathorne/vericlig>

³<http://www.nice.org.uk/nicemedia/pdf/CG87NICEGuideline.pdf>

- *activities* (e.g., providing advice, controlling blood glucose levels), representing units of execution in the process;
- participants, viz., the *actors* (e.g., doctors, nurses, patients), represented using pools, which are independent, autonomous points of execution, and possibly lanes, detailing participants belonging to the same pool;
- *artifacts* or *resources* (e.g., metmorfin) used or consumed by activities;
- *control flows and gates* (e.g., “if... then... else” control structures) that specify the acceptable orderings among activities inside a pool;
- *message flows*, representing information exchange between activities and participants belonging to different pools.

1.2 Process-evoking Words (PEWs)

In English, *content words* provide the vocabulary of the domain, denoting the objects, sets and (non-logical) relations that hold therein; their meaning (denotation) is static. On the other hand, *function words* denote the logical constraints, relationships and operations holding over such sets and relations. This distinction holds also to some degree (as allowed by their inherent ambiguity) in clinical domain documents, giving way to *process-evoking word categories* (and constituents).

Figure 1 provides an excerpt taken from a diabetes guideline. In it, activities, actors and artifacts/resources (i.e., static information) are denoted by content words. Activities are denoted often by transitive, intransitive or ditransitive verbs, viz., **VBs**⁴ and **VBZs**, participles (**VBNs**), gerunds (**VBGs**), etc, while actors and resources are denoted by the **NP** complements of such verbs. Control flows (i.e., dynamic, temporal information) are, on the other hand, denoted by function words, i.e., by *(i)* connectives introducing subordinated or coordinated phrases (e.g., **INs** such as “if”, in Figure 1), and *(ii)* temporal adverbs (e.g., “following”, in Figure 1) or prepositions (e.g., “after”, in Figure 1). Such connectives and adverbs are called in NLP literature *discourse relations* since they are used to combine together phrases (noun and verb phrases) and sentences (whether main or subordinated).

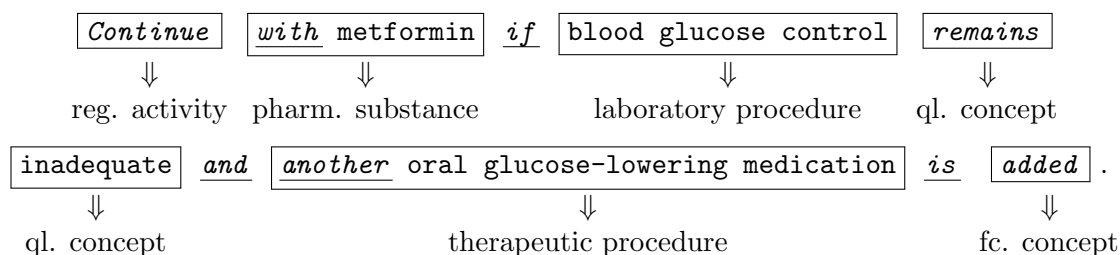
1.3 Negative Polarity-evoking Words (NEGs)

Another important issue is information polarity in clinical guidelines and CIGs. Guidelines generally state rule-like kind of information: actions or procedures to be executed by the clinical staff (e.g., “continue with metmorphin”) in the event that a given condition holds (e.g., “if blood glucose control remains inadequate”). The information they convey can be *positive* (i.e., an explicit statement of what should be done) or *negative* (i.e., an explicit statement of what should not be done, via rule consequents such as “the patient should not eat sugary foods”). Explicit negative information in natural language is conveyed by so-called *polarity contexts*, viz., syntactic constituents dominated (following the natural tree ordering in constituency parse trees) by, for instance, negation words (the “not” word of category *****) or negated modal verbs (i.e., **MD*s** such as “should not”, “can not”, “will not”). This issue is relevant, since the semantics of and the complexity of reasoning with formal process representations (such as CIGs) is much higher in the presence of explicit negative information [4].

2 Extraction of CIGs

The main challenge in CIG extraction consists in how to combine semantic annotation techniques, focusing on content words (i.e., on entities and events), with syntactic annotation tech-

⁴In what follows we refer to Penn Treebank word category and syntactic constituent tags, see [8].



(S (VP (VB *Continue*) (PP (IN with) (NP (NN metformin))))
 (SBAR (IN *if*)
 (S (S (NP (NN blood) (NN glucose) (NN control))
 (VP (VBZ remains)(ADJP (JJ inadequate))))))
 (CC *and*)
 (S (NP (DT another) (JJ oral) (JJ glucose-lowering)
 (NN medication))
 (VP (VBZ is) (VP (VBN added))))))

Legend
<i>activity evoking</i>
<i>resource evoking</i>
<i>control evoking</i>

Figure 2: **Top:** MetaMap UMLS (automated) annotation of item 1.5.1.3 from Figure 1. Word highlighting is ours. Entity segmentation (boxes) and annotations are MetaMap’s. **Bottom left:** Parse tree obtained with the Stanford parser. Word highlighting is ours.

niques capable of understanding the control flow structure conveyed by discourse relations, and information extraction methods dealing with clinical English ambiguity. In this section we provide an overview of the research challenges and of our proposed methodology to tackle them.

2.1 Limitations of Current Biomedical Resources

Research in biomedical NLP has yielded significant semantic annotation resources. Above all, the US National Library of Medicine’s Unified Medical Language System (UMLS) Metathesaurus⁵ and the annotated corpora, SemRep and annotation tools built upon it, MetaMap and SemRel, as described by [1]. MetaMap is used to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text. MetaMap uses a knowledge-intensive approach based on symbolic, NLP and computational linguistics techniques. Besides being applied for both IE and data-mining applications, MetaMap is one of the foundations of the US National Library of Medicine Medical Text Indexer (MTI) which is being used for both fully and semi automatic indexing of biomedical literature. Other resources that need to be mentioned are the semantically annotated CLEF corpus by [10], and Mayo Clinic’s Java API cTAKES (version 2.5), by [11].

Such resources, however, are still of limited use for the CIG extraction task. Gold-standard annotated guidelines are scarce for training and evaluation, and, if available, are not always in the public domain or might not support all biomedical IE tasks. Table 1 shows the main features of the mentioned biomedical resources.

Crucially, bio-medical/clinical IE systems often overlook process control structure or improperly understand clinical documents. Figure 2 illustrates a SemRel/MetaMap UMLS annotation of example 1.5.1.3 from Figure 1, with its associated phrase structure parse tree. As the reader

⁵<http://www.nlm.nih.gov/research/umls/>

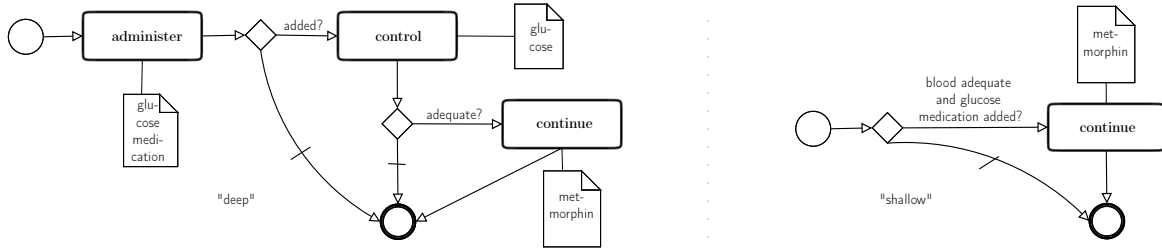


Figure 3: Two possible CIGs (in BPMN notation) of example 1.5.1.3 from Figure 1. Round boxes denote activities, diamonds conditional gates, square boxes resources and edges message flows. Circles represent the end (bold) and beginning of the process (normal), and barred edges “else” conditions.

Resource	ER	TE	RE	ANA	WSD	EV	Type	Free
CLEF	✓	✓	✓	✓	×	✓	Ann. corpus	×
UMLS	✓	✓	×	×	✓	×	Lexical resource	×
cTAKES 2.5	✓	✓	✓	✓	×	✓	Java API	✓
SemRep	✓	✓	✓	×	×	✓	Ann. corpus	×

Table 1: Main clinical and biomedical NLP resources and the features/IE tasks they support: entity recognition (ER), term extraction (TE), relation extraction (RE), anaphora resolution (ANA), word sense disambiguation (WSD) and event extraction (EV).

can see, such tools annotate only the identified “entities”, viz., the verbs and **NPs**, overlooking process structure as conveyed by discourse relations. Alternative IE techniques applied to guidelines such as [5] make extensive use of such resources. On the other hand, syntactic parsing, while necessary, as it can identify many discourse relations (e.g., the (**IN** *if*) constituent in Figure 2) and many of their arguments, is not sufficient, due to its limited domain knowledge: it provides little information regarding reified activities (e.g., the UMLS “procedures” in Figure 2). This may give rise to low precision, recall and accuracy for extraction methodologies based on either resource taken *alone*.

2.2 Business Process Extraction

The VERICLIG project seeks to understand whether these limitations can be overcome using techniques coming from the business processing community. [2] showed how to mine control flow semantics from parse trees (phrase structure and typed dependency trees) computed from business policy documents to extract (generic) business processes in BPMN notation with reasonably high accuracy (> 70%). We would like to adapt their general techniques to the clinical setting, by combining it with biomedical annotations.

The work by [2] has the advantage, moreover, of proposing alternative methods for measuring, e.g., extraction accuracy, less dependent on the availability of Gold corpora, by comparing the similarity of the extracted model with that of the workflow implemented in business information systems. As both workflows and process representations or models (in, e.g., BPMN notation) are embeddable in graphs, an appropriate evaluation metric is *graph edit distance*. Given the availability of careflows in clinical information systems, this evaluation strategy should be applicable to our case.

2.3 Clinical Word Sense Disambiguation

While function words and syntactic structure can prove informative to design rules to extract process control flows, domain knowledge is essential to extract the other components of CIGs, viz., activities, resources/artifacts and agents. As we argued before, and as the reader saw in Figure 2, such knowledge can be provided by the UMLS metathesaurus and its associated annotation tools. These however, do always necessarily (as in the example) assign a *unique* clinical interpretation/annotation to guideline content words (nouns and verbs, in particular), and when configured to do so, show low accuracy.

The problem of selecting or approximating, among the set of candidate (thesaurus) meanings, the intended interpretation of a word token within the context of the sentence, corpus and domain in which it occurs is known in NLP as the problem of *word sense disambiguation* (WSD). This problem has been studied extensively in NLP and computational linguistics (see [9]), but much less in the biomedical and clinical domains.

For instance, to assign to example 1.5.1.3 the “deep” CIG representation (in BPMN notation) from Figure 3, which we believe accurately captures its semantics, and not the “shallow” one, we need to *combine* the output of syntactic and semantic annotation techniques. Sem-Rep/MetaMap cannot extract process structure, but knows that “medication” and “control” (NNs) denote activities and not resources. Parsing, on the other hand, allows us to infer an “if...then...else” control structure, giving rise to the “shallow” process. However, by combining both sources of knowledge, we can see that the subordinated conditional phrase can be broken into a *sequence* of nested “if...then...else” structures.

As Table 1 shows, few (bio)medical resources support WSD. When they do (e.g., MetaMap), they do not do it robustly. Moreover, as we mentioned previously, CIG components such as, e.g., activities, are denoted not only by verbs, but also in many cases by nouns whose main denotation is not (as is the case with verbs) an event or activity, but a set. Every CIG extraction technique thus needs, crucially, to provide a clinical WSD methodology.

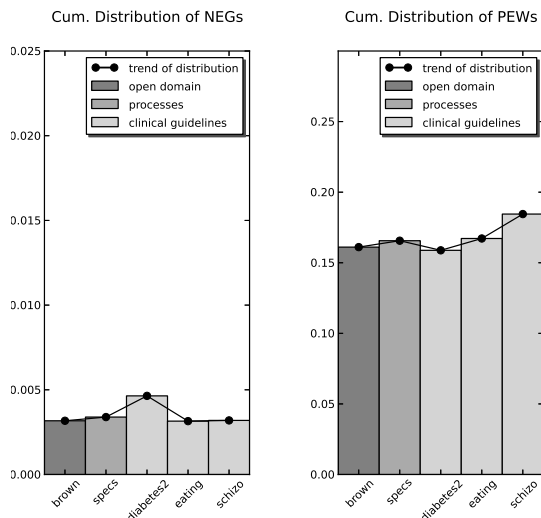
2.4 Methodology

We are currently developing a methodology based on business process extraction techniques and on WSD. In addition to this, we intend to provide in the future support for temporal relations and anaphora resolution as well, as anaphoric dependencies and temporal relations are needed to build complex models that interconnect the process fragments extracted from each of the guideline’s lines. As the reader may infer from Figure 3, process (and hence CIG) components temporally relate to each other, a feature spatially represented in BPMN’s graphical notation; thus, we also need to extract such relations. Our CIG extraction methodology can be summarized as follows:

1. combine annotation resources to extract CIG resources, actors and activities,
2. analyze syntactic/dependency structure to extract CIG control structure, and
3. resolve ambiguities and co-references, and infer temporal relations to build a complex CIG.

3 Guideline Analysis

In this section we study the distribution of PEWs and NEGs across different corpora. In particular, we want to know **(1)** if such words occur more frequently in closed-domain clinical and/or process corpora than in open-domain corpora such as the Brown corpus, or at any rate **(2)** if their frequency is correlated to corpus domain. By doing this, we can partially understand if PEWs and their word category or *part-of-speech* (POS) information can be used as control flow



Corpus	Size	Domain
Brown	1,391708 words	Open
Business	3,824 words	Business
Diabetes 2 guid.	7,109 words	Clinical
Eating dis. guid.	5,078 words	Clinical
Schizophr. guid.	5,367 words	Clinical

Figure 4: **Left.** Distribution of PEWs and NEG. **Right.** Corpora analyzed.

predictors, since (1) and (2) are necessary conditions for this to be the case. A negative answer for (1) or (2) means that richer models of process structure, exploiting complex syntactic and semantic features, are required. We analysed:

- a subset (A: press articles) of the Brown corpus⁶;
- Friederich’s corpus of business process specifications⁷;
- a subset (therapy recommendations) of the NICE diabetes-2 clinical guideline⁸;
- a subset (therapy recommendations) of the NICE eating disorders guideline⁹; and
- a subset (therapy recommendations) of the NCCN schizophrenia guideline¹⁰.

Figure 4 (right) summarizes their main features. They are all written for the most part in plain English. We used a pattern-based methodology that we describe below to collect the statistics. To answer questions (1) and (2), we reformulated the problem as an hypothesis testing/model verification problem in statistics, NLP and corpus linguistics [3, 7, 13]. In what follows we assume, moreover, that PEWs and NEG constitute *independent* events, thus disregarding their co-occurrence.

3.1 Pattern-based Analysis

We used two pattern-based methods, one for PEWs and another for NEG. In both cases we used first a POS tagger to annotate the raw corpora (the 3 guidelines and the specifications) and then checked for the occurrence of a certain number of target patterns. The patterns consisted a regular expression for each PEW or NEG, for their respective POSs, and for both. As before, we considered only Penn Treebank/Brown corpus POSs. For the POS tagging we relied on n-gram taggers: a 3-gram tagger, with 2-gram and unigram backoffs, trained over the (POS annotated) Brown corpus¹¹. Figure 4 (left) provides plots of the data collected.

⁶http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml

⁷<http://frapu.de/pdf/friedrich2010.pdf>

⁸<http://www.nice.org.uk/nicemedia/pdf/CG66NICEGuideline.pdf>

⁹<http://www.nice.org.uk/nicemedia/live/10932/29218/29218.pdf>

¹⁰<http://www.nice.org.uk/nicemedia/live/11786/43607/43607.pdf>

¹¹3-gram POS taggers (with backoffs) have, on average, > 80% accuracy.

Distribution of PEWs. PEWs are tokens belonging to the following POS categories:

- conjunctions and prepositions: **INs** (subordinating prepositions, e.g., “if”); **CCs** (coordinating conjunctions, e.g., “and”, “or”); **CCs** (subordinating conjunctions, e.g., “then”);
- adverbs: **RBs** (base adverbs, e.g., “after”); **RBRs** (comparative adverbs, e.g., “later”); **RBTs** (superlative adverbs, e.g., “latest”); **RNs** (nominalized adverbs, e.g.,); and **RPs** (adverbial particles, e.g.,).

Distribution of NEGs. NEGs, on the other hand, are tokens of the following POS categories:

- “not”, of category ***** (i.e., negation); “nobody”, “none” and “nothing”, of category **PN**, the negative determiner “no” of category **AT**; and
- negated modal verbs of category **MD*** (e.g., “cannot”, “will not”).

3.2 Hypothesis Testing

In the hypothesis testing setting, we typically test a so-called *null hypothesis* H_0 , that states that the pattern(s) observed in the data are random, against an *alternative hypothesis* H_1 , that encodes the model we think may fit the sample and generalize to the whole population. The test consists, briefly, in checking there is *significant evidence to reject* H_0 and *accept* H_1 , and proceeds by computing a test *score* or statistic τ that estimates how well the sample “fits” the null hypothesis.

More crucially, by comparing the τ score with a so-called τ -distribution, with possibly k degrees of freedom, it assigns to the score a *p-value*, that can be intuitively understood as the probability of the sample being not significant enough to reject H_0 . If such value, however, is sufficiently small, and if in particular is *smaller than* the *significance levels* (a.k.a. the *power* of the test) of either 0.05 (standard significance), 0.01 (strong significance) or 0.001 (very strong significance), rejection of H_0 and acceptance of H_1 ensues. For the detailed account of hypothesis testing in statistics and corpus linguistics we send the reader to [13].

We considered two test statistics and methods. To check **(1)**, we made use of a one-way/sample *t*-test, valid when noise and sampling error are normally distributed [] (which we assume to be the case). To check for **(2)**, we used a χ^2 -test for independence, that checks for absences of correlations and that has the advantage of being valid for possibly non-normally distributed data (as is the case for NL corpora [3, 7]).

The one-way/sample *t*-test. Student’s one-sample *t*-test is typically used to test the null hypothesis that $m = \mu_0$, viz., that the sample’s mean m is equal to the expected mean μ_0 , with differences across samples due only to chance (and distributed normally). It is obtained by computing the so-called *t-score*:

$$t = \frac{m - \mu_0}{s/\sqrt{n}} \quad (t\text{-score})$$

where m and μ_0 are as before, s is the sample’s standard deviation and n its size (the number of observations), and comparing it to the *t*-distribution with $n - 1$ degrees of freedom.

We used the *t*-test to check if the observed frequency distributions differ across corpora only by chance, or vary normally independently of corpus type. We tested:

1. for PEWs the null hypothesis H_0 that $m = \mu_0 = 0.25$, under 4 degrees of freedom, with a $p < 0.01$ significance level, against the alternative hypothesis H_1 that the distribution of PEWs is skewed towards (clinical) process documents; and

Distrib.	χ^2 -ind.	p (< 0.001 sig.)	df.	t -one way	μ_0	p (< 0.01 sig.)	df.
PEWs	9.39	0.009	2	1.02	0.20	0.36	4
NEGs	1.96	0.375	2	1.02	0.03	0.36	4

Table 2: Statistical tests.

- for the NEGs, the null hypothesis H'_0 that $m = \mu_0 = 0.05$, under 4 degrees of freedom, with, again, a $p < 0.01$ significance level, against the alternative hypothesis H'_1 that distribution of NEGs is skewed towards (clinical) process documents.

The χ^2 -test of independence. The χ^2 -test of independence, tests the null hypothesis that two categorical variables X and Y are independent, viz., that no correlation exists among them. It is obtained by computing the χ^2 -score

$$\chi^2 = \sum_{i,j} (O_{i,j} - E_{i,j})^2 / E_{i,j} \quad (\chi^2\text{-score})$$

where $O_{i,j}$ and $E_{i,j}$ are, resp., the observed and expected value (the raw frequency or count) for (the joint) outcome x_i of variable X of dimension n and outcome y_j of variable Y of dimension m , arranged in a $n \times m$ table, and comparing it to the χ^2 -square distribution with $(n-1) \times (m-1)$ degrees of freedom.

We considered as variables: (i) domain D , with 3 outcomes (i.e., the corpus domains), and (ii) word type T and T' , with 2 outcomes each (i.e., whether a word is a PEW or a NEG, or whether it is not, resp.), and tested

- for PEWs the null hypothesis H_0 that D and T are independent, under 1×2 degrees of freedom and $p < 0.001$ significance, against the hypothesis H_1 that the distribution of PEWs is in correlation with document domain (and skewed towards the clinical, in particular); and
- for the NEGs, the null hypothesis H'_0 that D and T' are independent, under 1×2 degrees of freedom and $p < 0.001$ significance, against the hypothesis H'_1 that the distribution of NEGs is correlated to the clinical domain.

3.3 Results and Discussion.

Table 2 summarizes the results of both tests. As the reader can see, in both cases (the t -test and the χ^2 -test for independence) the tests showed that there is no significant change in PEW/NEG distribution/frequency across corpora and no significant correlation between PEW/NEG distribution/frequency and corpus type. Thus, it may seem that any perceived bias towards the clinical therapy and process domain is due only to chance, and that questions (1) and (2) should receive a negative answer. They seem rather to follow their distribution in ordinary English, even if at a level of 0.05 significance, one may conclude that PEW distribution seems significantly correlated/skewed towards clinical guideline documents.

This may seem to imply¹² that words and word categories only very roughly approximate the control flows or negations intended by guidelines, and that, taken alone, are bad predictors. Denotation or semantic interpretation in English (and NL) is, in general, a *many-to-many* relation, due to semantic phenomena such as ambiguity, polysemy or homonymy. Therefore, in the pattern-based methodology for CIG extraction described earlier, rich models of (lexical, syntactic or semantic) context will be necessary to extract the control flow components of CIGs.

¹²To the extent that the corpora considered in our study are deemed representative.

Furthermore, as the reader can see in Figure 4 (left), the (mean) relative frequency of PEWs is very low overall (≤ 0.005), and that hence explicit negative information in processes (i.e., beyond a CWA “negation as failure” setting) may seem irrelevant for CIGs.

4 Results for 2012 and Work Plan for 2013

4.1 Results for 2012

We have collected a corpus of guidelines related to chronic diseases (e.g., diabetes, obesity, food allergy, etc.)¹³, and plan to collaborate with the Merano hospital in Merano, Italy and with the FBK’s eHealth group in Trento, Italy, to acquire the required careflows/clinical workflows for CIG extraction evaluation. This collaboration has produced a position paper (see [14]) describing the VERICLIG project, that will be presented at the 2013 Workshop on Computational Semantics for Clinical Text (ClinText2013), at Potsdam, this month of March.

4.2 Plan for 2013

In the near future, we intend to follow the next steps: **(1)** We plan to implement a baseline system to evaluate our proposed methodology along the lines discussed above. **(2)** In particular, we plan to investigate WSD methodologies for activity identification, as a way of tackling the CIG/clinical process “granularity” problem. **(3)** To further refine CIG extraction, we will also consider applying formal methods (e.g., temporal logic reasoning) to prune logically inconsistent CIGs and/or their components. **(4)** We also plan to write and submit a survey paper on CIG extraction to a clinical informatic journal (e.g., BMC Bioinformatics), given the lack of literature on the subject.

References

- [1] Alan R. Aronson and François-Michel Lang. An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.
- [2] Fabian Friederich, Jan Mendling, and Frank Puhmann. Process model generation from natural language text. In *Proc. of the 23rd Int. Conf. on Advanced Information Systems Engineering (CAiSE 2011)*, 2011.
- [3] 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.
- [4] Arjen Hommenrsom, Perry Groot, Michael Balser, and Peter Lucas. Formal methods for verification of clinical practice guidelines. In A. Ten Teije et al., editor, *Computer-based Medical Guidelines and Protocols: A Primer and Current Trends*, chapter 4, pages 63–80. IOS Press, 2008.
- [5] Katharina Kaiser, Cem Akkaya, and Silvia Miksch. Gaining process information from clinical practice guidelines using information extraction. In *Proc. of the 10th Int. Conf. on Artificial Intelligence on Medicine (AIME 2005)*, 2005.
- [6] Ryan K.L. Ko, Stephen S.G. Lee, and Eng Wah Lee. Business process management (BPM) standards: A survey. *Business Process Management J.*, 15(5):744–791, 2009.
- [7] Christopher Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 2000.

¹³Available from the authors by request.

- [8] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.
- [9] Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69, 2009.
- [10] Angus Roberts, Robert Gaizaskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, and Bill Wheeldin. The CLEF corpus: Semantic annotation of a clinical text. In *Proc. of the AMIA 2007 Annual Symp.*, 2007.
- [11] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 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.
- [12] Yuval Shahaar, 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.
- [13] Tsu T. Soong. *Fundamentals of Probability and Statistics for Engineers*. Wiley, 2004.
- [14] Camilo Thorne, Elena Cardillo, Marco Montali, Caludio Eccher, and Diego Calvanese. The VERICLIG project: Extraction of computer interpretable guidelines via syntactic and semantic annotation. In *Proceedings of the 2013 Workshop on Computational Semantics for Clinical Text (ClinText 2013)*, 2013.