Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis

S. Velupillai1, D. Mowery2, B. R. South2, M. Kvist1,3, H. Dalianis1
1 Department of Computer and Systems Sciences, (DSV), Stockholm University, Stockholm, Sweden
2 Department of Biomedical Informatics, University of Utah, Salt Lake City, USA
3 Department of Learning, Informatics, Management and Ethics (LIME), Karolinska Institutet, Sweden

Summary

Objectives: We present a review of recent advances in clinical Natural Language Processing (NLP), with a focus on semantic analysis and key subtasks that support such analysis.

Methods: We conducted a literature review of clinical NLP research from 2008 to 2014, emphasizing recent publications (2012-2014), based on PubMed and ACL proceedings as well as relevant referenced publications from the included papers.

Results: Significant articles published within this time-span were included and are discussed from the perspective of semantic analysis. Three key clinical NLP subtasks that enable such analysis were identified: 1) developing more efficient methods for corpus creation (annotation and de-identification), 2) generating building blocks for extracting meaning (morphological, syntactic, and semantic subtasks), and 3) leveraging NLP for clinical utility (NLP applications and infrastructure for clinical use cases). Finally, we provide a reflection upon most recent developments and potential areas of future NLP development and applications.

Conclusions: There has been an increase of advances within key NLP subtasks that support semantic analysis. Performance of NLP semantic analysis is, in many cases, closer to that of agreement between humans. The creation and release of corpora annotated with complex semantic information models has greatly supported the development of new tools and approaches. Research on non-English languages is continuously growing. NLP methods have sometimes been successfully employed in real-world clinical tasks. However, there is still a gap between the development of advanced resources and their utilization in clinical settings. A plethora of new clinical use cases are emerging due to established health care initiatives and additional patient-generated sources through the extensive use of social media and other devices.

Keywords
Clinical Natural Language Processing; Semantics; Information Extraction; Annotation; Domain Adaptation; Review

Introduction

Communication is a fundamental process that supports our daily existence. We represent information and knowledge by using natural language to describe entities and events and their relationships to each other and across time. The concepts underlying the linguistic expressions convey meaning, or semantics, of how entities (people, places, or things) interact with each other. In today’s world, we use a variety of text types - emails, blogs, SMS texts, reports, etc. - to facilitate communication about these interactions. The meaning conveyed in communication is only available for human consumption and not for machine interpretation unless we can map the text to the underlying semantics. Natural Language Processing (NLP) is an approach for automatically encoding the semantics represented in natural language texts.

Clinical NLP is the application of text processing approaches on documents written by healthcare professionals in clinical settings, such as notes and reports in health records. The interest for clinical NLP is spurred by the need for real-time, large-scale, and accurate information extraction from health records to support clinical care, e.g., through automated generation of a patient problem list, to support biomedical and health services research, e.g., through precise cohort identification, and to support public health practice, e.g., through disease surveillance. Clinical NLP can provide clinicians with critical patient case details, which are often locked within unstructured clinical texts and dispersed throughout a patient’s health record. Semantic analysis is one of the main goals of clinical NLP research and involves unlocking the meaning of these texts by identifying clinical entities (e.g., patients, clinicians) and events (e.g., diseases, treatments) and by representing relationships among them.

The most crucial step to enable semantic analysis in clinical NLP is to ensure that there is a well-defined underlying schematic model and a reliably-annotated corpus, that enables system development and evaluation. It is also essential to ensure that the created corpus complies with ethical regulations and does not reveal any identifiable information about patients, i.e. de-identifying the corpus, so that it can be more easily distributed for research purposes.

Once these issues are addressed, semantic analysis can be used to extract concepts that contribute to our understanding of patient longitudinal care. For example, lexical and conceptual semantics can be applied to encode morphological aspects of words and syntactic aspects of phrases to represent the meaning of words in texts. However, clinical texts can be laden with medical jargon and can be composed with telegraphic constructions. Furthermore, sublanguages can exist within each of the various clinical sub-domains and note types [1-3]. Therefore, when applying computational semantics, automatic processing of semantic meaning from texts, domain-specific methods and linguistic features for accurate parsing and information extraction should be considered.

In clinical practice, there is a growing curiosity and demand for NLP applications. Today, some hospitals have in-house solutions or legacy health record systems for which NLP algorithms are not easily applied. However, when applicable, NLP could play an important role in reaching the goals of better clinical and population health outcomes by the improved use of the textual content contained in EHR systems.
In this paper, we review the state of the art of clinical NLP to support semantic analysis for the genre of clinical texts.

Methods

We present a review of clinical NLP research that supports semantic analysis of clinical texts, with an emphasis on recent publications (2012-2014), but also including relevant work published after 2010. Selection criteria are described in Zweigenbaum and Névéol [7]. For a related survey on articles published earlier, we refer to Meystre et al. [4]. Our scope limitations restrict in-depth coverage of research in closely related areas. For instance, although similar in medical and scientific topics, biomedical texts, patient-generated documentation, and forum discussions are not included in this review. For recent advances also in biomedical NLP, we refer to Cohen & Demner-Fushman [8].

We will structure our review by addressing three main tasks for applying semantic analysis to clinical texts: 1) developing more efficient methods for corpus creation, 2) generating linguistic and semantic building blocks for extracting meaning, and 3) applying NLP applications to clinical use cases. More specifically, we will review recent clinical NLP research that has addressed the aforementioned barriers to progress including the development of more efficient methods for corpus creation in the context of annotation and de-identification. We will then elaborate upon notable developments in generating the building blocks of semantic analysis addressing subtasks of morphological and syntactic processing as well as semantic processing subtasks of named entity recognition and contextual analysis, co-reference resolution, temporal reasoning, and document-level analysis. Further, we will describe efforts toward developing NLP applications and systems that demonstrate clinical utility of NLP leveraging semantic processing for clinical use cases. Use cases addressed include identifying disease and associated symptomology, assigning diagnostic billing codes, detecting adverse drug events, and monitoring for hospital-acquired infections. We will conclude with a summary of the state of the art and discuss potential up-and-coming frontiers for clinical NLP and semantic analysis in the near future.

In the following sections, we describe the current state of the art in clinical NLP towards the support of semantic analysis within these three thematic areas: corpus creation, semantic building blocks, and clinical utility.

Corpus Creation - Supporting Semantic Analysis with Efficiency and Accessibility

Two of the most important first steps to enable semantic analysis of a clinical use case are the creation of a corpus of relevant clinical texts, and the annotation of that corpus with the semantic information of interest. For example, if we are interested in developing an NLP system that helps maintain accurate problem lists for patients in intensive care units (ICU), one could obtain clinical notes from retrospective ICU patient records and encode common problems of interest, e.g., signs, symptoms, and diagnoses, for developing and evaluating an NLP system for that use case. Identifying the appropriate corpus and defining a representative, expressive, unambiguous semantic representation (schema) is critical for addressing each clinical use case.

Annotation — Developing Reliable and Sufficient Datasets

Once a corpus is selected and a schema is defined, it is assessed for reliability and validity [9], traditionally through an annotation study in which annotators, e.g., domain experts and linguists, apply or annotate the schema on a corpus. Ensuring reliability and validity is often done by having (at least) two annotators independently annotating a schema, discrepancies being resolved through adjudication. Pustejovsky and Stubbs present a full review of annotation designs for developing corpora [10]. However, manual annotation is time consuming, expensive, and labor intensive on the part of human annotators. Methods for creating annotated corpora more efficiently have been proposed in recent years, addressing efficiency issues such as affordability and scalability. With such methods, advancements in semantic analysis are enabled.

Affordability

One major barrier to corpus annotation is affordability. Minimizing the manual effort required and time spent to generate annota-
Scalability can be defined by aspects of resources needed to develop a reliable and valid reference standard including the type of annotator expertise, the number of annotators, and the number of texts. For instance, more experienced annotators can command higher pay. Employing more annotators and annotating a larger corpus than what is needed for high reliability and validity can result in greater, unnecessary costs. Recent efforts leveraging crowdsourcing technologies assess the training of a crowd of non-domain experts rather than a set of domain experts to create a large and reliable reference standard quickly. Zhai et al. [16] built a reference standard of medication annotations for clinical trial announcements from the ClinicalTrials.gov website using CrowdFlower, an Amazon Mechanical Turk-based crowdsourcing platform, resulting in high human agreement (>73% F1) and showing that there was no statistically significant difference between crowd- and expert-generated annotations for this task. Similarly, in the biomedical NLP community, crowdsourcing has been used to produce annotated data successfully [17]. Although this solution is difficult to apply on clinical texts due to confidentiality reasons, the clinical NLP community could still benefit from these experiences to make headway. For instance, in the 2009 i2b2 challenge on medication extraction [18], participating teams were required to also produce annotations, thus minimizing costs for the challenge organizers to hire external annotators.

Other efforts systematically analyzed what resources, texts, and pre-processing are needed for corpus creation. Jucket [19] proposed a generalizable method using probability weighting to determine how many texts are needed to create a reference standard. The method was evaluated on a corpus of dictation letters from the Michigan Pain Consultant clinics. Gundlapalli et al. [20] assessed the usefulness of pre-processing by applying v3NLP, a UIMA-AS-based framework, on the entire Veterans Affairs (VA) data repository, to reduce the review of texts containing social determinants of health, with a focus on homelessness. Specifically, they studied which note titles had the highest yield (‘hit rate’) for extracting psychosocial concepts per document, and of those, which resulted in high precision. This approach resulted in an overall precision for all concept categories of 80% on a high-yield set of note titles. They conclude that it is not necessary to involve an entire document corpus for phenotyping using NLP, and that semantic attributes such as negation and context are the main source of false positives.
new corpora for developing and evaluating de-identification applications.

Several systems and studies have also attempted to improve PHI identification while addressing processing challenges such as utility, generalizability, scalability, and inference.

Utility
Utility of clinical texts can be affected when clinical eponyms such as disease names, treatments, and tests are spuriously redacted, thus reducing the sensitivity of semantic queries for a given use case. For example, if mentions of Huntington's disease are spuriously redacted from a corpus to understand treatment efficacy in Huntington's patients, knowledge may not be gained because disease/treatment concepts and their causal relationships are not extracted accurately. One de-identification application that integrates both machine learning (Support Vector Machines (SVM), and Conditional Random Fields (CRF)) and lexical pattern matching (lexical variant generation and regular expressions) is BoB (Best-of-Breed) [25-26]. BoB applies the highest performing approaches from known de-identification systems for each PHI type, resulting in balanced recall and precision results (89%) for a configuration of individual classifiers, and best precision (95%) was obtained with a multi-class configuration. This system was also evaluated to understand the utility of texts by quantifying clinical information loss following PHI tagging i.e., medical concepts from the 2010 i2b2 Challenge corpus, in which less than 2% of the corpus concepts partially overlapped with the system [27].

Generalizability
Generalizability is a challenge when creating systems based on machine learning. In particular, systems trained and tested on the same document type often yield better performance, but document type information is not always readily available. By creating training and testing sets on clinical documents that were partitioned into similar types – categorized by measures of writing complexity and clinical vocabulary usage – de-identification results were improved (avg. F1 92%), compared to using randomly selected clusters (avg. F1 88%), on a collection of over 4500 various document types from Vanderbilt University Medical Center [28].

Another challenge related to generalizability is to apply and evaluate de-identification methods not only on various document types, but also on corpora in non-English languages. Toward this goal, Grouin and Névéol [29] developed a reference corpus in French for de-identification. Two pre-annotation tools were used, one rule-based and one CRF-based, evaluated by two annotators. The rule-based system produced better pre-annotations, but the manual revision of the CRF-based system was faster. In both cases, human agreement was high (> 90% F1), and only 20 documents were needed to build a statistical system that outperformed pre-annotation tools. Two gold standard sets of French clinical notes were created. For Swedish, CRF models were also used to refine a set of de-identification annotations, along with manual revision, resulting in a new gold standard and system performance of 80% F1 [30]. PHI annotations on a subset of this corpus (100 notes) have been pseudonymized [31] and approved for release to the research community. In Korea, clinical notes are written both in English and Korean. To handle Korean names in a de-identification system, a heuristic approach using regular expressions was adopted and verified on 6,502 clinical notes from the Asian Medical Center, resulting in 89% precision and 97% recall [23].

Scalability
Scalability of de-identification for larger corpora is also a critical challenge to address as the scientific community shifts its focus toward “big data”. Deleger et al. [32] showed that automated de-identification models perform at least as well as human annotators, and also scales well on millions of texts. This study was based on a large and diverse set of clinical notes, where CRF models together with post-processing rules performed best (93% recall, 96% precision). Moreover, they showed that the task of extracting medication names on de-identified data did not decrease performance compared with non-anonymized data.

Inference
Inference that supports semantic utility of texts while protecting patient privacy is perhaps one of the most difficult challenges in clinical NLP. Privacy protection regulations that aim to ensure confidentiality pertain to a different type of information that can, for instance, be the cause of discrimination (such as HIV status, drug or alcohol abuse) and is required to be redacted before data release. This type of information is inherently semantically complex, as semantic inference can reveal a lot about the redacted information (e.g., *The patient suffers from XXX (AIDS) that was transmitted because of an unprotected sexual intercourse*). Sánchez et al. [33] describe a method to sanitize clinical texts without disclosure from semantic inference using information theoretic measures, knowledge bases, and the Web as corpora with promising results when evaluated on Wikipedia descriptions of medical entities considered as sensitive by United States state and federal laws.

Semantic Building Blocks
— Extracting Meaning From Texts

Semantic analysis can be a powerful tool for representing information and conveying meaning from clinical texts. Clinical NLP pipelines apply semantic analysis of clinical texts by integrating several meta-layers of textual information into standard information models. These information models not only describe semantic concepts, their attributes, and their interactions and relations with each other to convey meaning, but also how linguistic information, such as syntax, can be used to accurately fill these arguments and relations within a semantic structure. For example, syntactic and semantic information can represent events experienced by a person using a frame structure, e.g., “patient underwent chemotherapy” can be represented as:

S \(\rightarrow\) Arg1[patient/SBJ NP (experiencer)],
Rel[undergo.1/VP(experience, undergo)],
Arg2[chemotherapy/OBJ NP (experienced)]
Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis

Several types of textual or linguistic information layers and processing—morphological, syntactic, and semantic—can support semantic analysis.

**Morphological and Syntactic Processing—Encoding Linguistic Layers for Semantics**

Morphological and syntactic preprocessing can be a useful step for subsequent semantic analysis. For example, prefixes in English can signify the negation of a concept, e.g., *afebrile* means *without fever*. Furthermore, a concept’s meaning can depend on its part of speech (POS), e.g., discharge as a noun can mean *fluid from a wound*; whereas a verb can mean *to permit someone to vacate a care facility*. Many of the most recent efforts in this area have addressed adaptability and portability of standards, applications, and approaches from the general domain to the clinical domain or from one language to another language.

**Adaptability to the Clinical Domain**

Several standards and corpora that exist in the general domain, e.g., the Brown Corpus and Penn Treebank tag sets for POS-tagging, have been adapted for the clinical domain. Fan et al. [34] adapted the Penn Treebank II guidelines [35] for annotating clinical sentences from the 2010 i2B2/VA challenge notes with high inter-annotator agreement (93% F1). This adaptation resulted in the discovery of clinical-specific linguistic features. This new knowledge was used to train the general-purpose Stanford statistical parser, resulting in higher accuracy than models trained solely on general or clinical sentences (81%).

New morphological and syntactic processing applications have been developed for clinical texts. cTAKES [36] is a UIMA-based NLP software providing modules for several clinical NLP processing steps, such as tokenization, POS-tagging, dependency parsing, and semantic processing, and continues to be widely-adopted and extended by the clinical NLP community. The variety of clinical note types requires domain adaptation approaches even within the clinical domain. One approach called ClinAdapt uses a transformation-based learner to change tag errors along with a lexicon generator, increasing performance by 6-11% on clinical texts [37].

**Portability to New Languages**

A statistical parser originally developed for German was applied on Finnish nursing notes [38]. The parser was trained on a corpus of general Finnish as well as on small subsets of nursing notes. Best performance was reached when trained on the small clinical subsets than when trained on the larger, non-domain specific corpus (Labeled Attachment Score 77-85%). To identify pathological findings in German radiology reports, a semantic context-free grammar was developed, introducing a vocabulary acquisition step to handle incomplete terminology, resulting in 74% recall [39].

**Semantic Processing — Representing Meaning from Texts**

To fully represent meaning from texts, several additional layers of information can be useful. Such layers can be complex and comprehensive, or focused on specific semantic problems. In recent years, several efforts have addressed semantic processing subtasks from the perspective of *information models and shareable resources*—an instrumental part for semantic analysis method development, in areas such as *named entity recognition and contextual attributes, coreference resolution, temporal reasoning, and document-level analysis*.

**Semantic Analysis Method Development — Information Models and Resources**

One notable effort for a rich information model with several annotation layers is the MiPACQ (Multi-source Integrated Platform for Answering Clinical Questions) dataset consisting of annotated Treebank POS tagged tokens, PropBank predicate-argument frames, and UMLS encoded entities on a corpus of randomly-selected Mayo Clinic clinical and pathology notes related to colon cancer with high agreement (93%, 89-93% and 70-75%, respectively) [40]. This dataset has promoted the dissemination of adapted guidelines and the development of several open-source modules.

Other development efforts are more dependent on the integration of several information layers that correspond with existing standards. The latter approach was explored in great detail in Wu et al. [41] and resulted in the implementation of the secondary use Clinical Element Model (CEM) [42] with UIMA, and fully integrated in cTAKES [36] v2.0.

The organization of shared tasks, or community challenges, has also been an influential part of the recent advancements in clinical NLP not only in corpus creation and release, annotation guideline development and schema modeling, but also in defining semantically-related tasks. Furthermore, NLP method development has been enabled by the release of these corpora, producing state-of-the-art results [17].

Many of these corpora address the following important subtasks of semantic analysis on clinical text.

**Named Entity Recognition and Contextual Analysis**

Correctly identifying the entities or concepts to which semantic modifiers or relations belong is crucial for information extraction. Often, concepts are defined as noun phrases (e.g. *diabetes mellitus*), requiring algorithm solutions that deal with sequences of words, either rule- or machine learning-based, as shown in the solutions for previous challenges—the 2010 i2b2 challenge [43], the 2014 ShARe/CLEF eHealth challenge [44] and the SemEval 2015 Task 14: Analysis of Clinical Text [45]—on concept classification, with system performance as high as 85% F1. In clinical settings, semantic type information is also essential, for instance, knowing that a concept is a problem, a test or a treatment, as in the definition of the 2010 i2b2 challenge.
Clinical entity recognition has also been studied for non-English languages. For instance, one study employed CRF models on Swedish clinical data for the types disorders, findings, pharmaceuticals and body structures, resulting in F1s ranging between 69-81% [46], in line with or slightly lower than results reported for English.

For accurate information extraction, contextual analysis is also crucial, particularly for including or excluding patient cases from semantic queries, e.g., including only patients with a family history of breast cancer for further study. Contextual modifiers include distinguishing asserted concepts (patient suffered a heart attack) from negated (not a heart attack) or speculative (possibly a heart attack). Other contextual aspects are equally important, such as severity (mild vs severe heart attack) or subject (patient or relative).

The ShARe (Shared Annotated Resources) corpus - a subset of discharge summaries, radiology, echocardiogram, and electrocardiogram reports from the MIMIC II database [47] - consists of templates with disease/disorder events encoded with SNOMED CT concept unique identifiers (CUI), with rich contextual attributes from the Clinical Element Model (CEM) [42] and with temporal expression mentions [44]. This dataset is unique in its integration of existing semantic models from both the general and clinical NLP communities.

In the 2014 ShARe/CLEF eHealth task 2, in an effort to leverage this annotated dataset, several approaches were taken to normalize semantic modifiers such as body site and severity which included optimizing cTAKES modules, developing rules based on resources from UMLS, and employing grammatical relations [48]. For example, Diligach et al. [49] treat these two problems as a relation extraction task, building SVM models and evaluating on two clinical corpora, resulting in F1 scores of 74-91% for body site and 91-93% for severity. The created models have been released as cTAKES modules. In addition to normalization of specific modifiers, the SemEval 2015 Task 14: Analysis of Clinical Texts also included an end-to-end system evaluation [45] that assessed NLP system performance for identifying and normalizing disease/disorders and their modifiers from the ShARe corpus.

An ensemble machine learning approach leveraging MetaMap and word embeddings from unlabeled data for disorder identification, a vector space model for disorder normalization, and SVM approaches for modifier classification achieved the highest performance (combined F1 and weighted accuracy of 81%) [50].

When encoding semantic concepts, lexicon- and rule-based NLP systems have the advantage of being almost language independent since the underlying algorithms do not necessarily depend on the source language. However, they require language-specific rules and lexicons. pyConTextNLP [51], a rule-based system for classifying assertions (negation and/or uncertainty modifiers) of disease mentions, was ported from English to Swedish [52]. The system relies on a cue lexicon with scoping rules. To create a Swedish lexicon, the authors translated and added cues from several sources, and final overall results were reported as 81% F1. When comparing negation and uncertainty cues across languages, the most frequent cues are often similar, but rarer cues are more prone to individual language particularities [53-54]. Further, the negation detection algorithm NegEx was evaluated on additional languages, where negation cues were translated also to French and German [53]. The system has also been adapted to Dutch, where adaptations for the contextual attributes negation, experience, and temporality were developed through translations from English along with enhanced rules and regular expressions [55]. Final results for negation and experience were high (> 87% and > 99% F1, respectively), but lower for historical and hypothetical temporality properties (26-54% and 13-44% F1, respectively).

Experience and temporality attributes were also studied as a classification task on a corpus of History and Physical Examination reports, where the ConText algorithm was compared to three machine learning (ML) algorithms (Naïve Bayes, k-Nearest Neighbours and Random Forest). There were no statistically significant differences in results for classifying expercer between these approaches, but the ML approach (specifically, Random Forest) outperformed ConText on classifying temporality (historical or recent), resulting in 87% F1 compared to 69% [56].

Coreference Resolution

A challenging issue related to concept detection and classification is coreference resolution, e.g., correctly identifying that it refers to heart attack in the example “She suffered from a heart attack two years ago. It was severe.” NLP approaches applied on the 2011 i2b2 challenge corpus included using external knowledge sources and document structure features to augment machine learning or rule-based approaches [57]. For instance, the MCORES system employs a rich feature set with a decision tree algorithm, outperforming unweighted average F1 results compared to existing open-domain systems on the semantic types Test (84%), Persons (84%), Problems (85%) and Treatments (89%) [58]. Another approach deals with the problem of unbalanced data and defines a number of linguistically and semantically motivated constraints, along with techniques to filter co-reference pairs, resulting in an unweighted average F1 of 89% [59]. Domain knowledge and domain-inspired discourse models were employed by Jindal & Roth on the same task and corpus with comparable results (unweighted average F1 between 84-88%), where the authors concluded that most recall errors could be handled by addition of further domain knowledge [60].

Temporal Reasoning

Temporal modeling has been the focus of many recent studies. In order to understand disease progression, adverse drug reactions, and other clinically relevant events over time, semantic models of temporality are needed. For instance, knowing when particular symptoms were present for a specific disease can be used to build predictive models to ensure timely treatment. The TimeML model [61], a rich model to represent temporal information in text through events (high blood pressure, heart attack), time expressions (two years ago), and their temporal relation (high blood pressure BEFORE heart attack), has been adapted in at least two ways to the clinical domain. Styler et al. [62] adapted the model to pathology and clinical texts from Mayo clinic, creating the THYME (Temporal History of Your Medical Events) corpus, while another adaptation was used in the 2012 i2b2
Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis

Clinical Utility — Applying NLP Applications to Clinical Use Cases

In order to employ NLP methods for actual clinical use-cases, several factors need to be taken into consideration. Many (deep) semantic methods are complex and not easy to integrate in clinical studies, and, if they are to be used in practical settings, need to work in real-time. Several recent studies with more clinically-oriented use cases show that NLP methods indeed play a crucial part for research progress. Often, these tasks are on a high semantic level, e.g. finding relevant documents for a specific clinical problem, or identifying patient cohorts. For instance, NLP methods were used to predict whether or not epilepsy patients were potential candidates for neurosurgery [80]. Clinical NLP has also been used in studies trying to generate or ascertain certain hypotheses by exploring large EHR corpora [81]. In other cases, NLP is part of a grander scheme dealing with problems that require competence from several areas, e.g. when connecting genes to reported patient phenotypes extracted from EHRs [82-83].

Identifying Disease and Associated Symptomatology

A method for identifying progress notes pertaining to diabetes was developed using a supervised machine learning framework (SVM), using a bag-of-words (BoW) representation, on notes from different institutions, resulting in F1 scores > 93% [84]. Interestingly, richer NLP features (named entities, synonym resolution, negation) were not found to be useful for this task. Similarly, Yetisgen-Yildiz et al. [85] obtained best results (79% F1) using n-gram features (uni-, bi- and trigrams) in a study to identify patients confirmed to have Acute Lung Injury. They found that assertion values were not useful for this task. Their study was evaluated on a corpus of 1,748 free-text chest x-ray reports related to patients at an intensive care unit at the Harborview Medical Center.

Another example in psychiatry showed that models incorporating NLP (using the HiTeX...
Assigning Diagnostic Billing Codes
ICD-9 and ICD-10 (version 9 and 10 respectively) denote the international classification of diseases [89]. ICD codes are usually assigned manually either by the physician herself or by trained manual coders. They are used primarily for billing purposes for hospital administrations. Manually assigned codes are, however, often erroneous. In an investigation carried out by the National Board of Health and Welfare (Socialstyrelsen) in Sweden, 4,200 patient records and their ICD-10 coding were reviewed, and they found a 20 percent error rate in the assignment of main diagnoses [90]. NLP approaches have been developed to support this task, also called automatic coding, see Stanfill et al. [91], for a thorough overview. The best performing systems obtain F1-scores of around 90%. Perotte et al. [92], elaborate on different metrics used to evaluate automatic coding systems. Other recent approaches for automatic coding support are described in e.g. Martinez et al. [93].

Detecting Adverse Drug Events
An important aspect in improving patient care and healthcare processes is to better handle cases of adverse events (AE) and medication errors (ME). A study where NLP was used to automate detection of IV infiltrations, narcotic medication over sedation and dosing errors in a neonatal intensive care setting provided higher sensitivity and positive predicted value compared with manual trigger tools (as high as 100% precision and recall for some AE/ME types) [94]. A study on Danish psychiatric hospital patient records [95] describes a rule- and dictionary-based approach to detect adverse drug effects (ADEs), resulting in 89% precision, and 75% recall. Another notable work reports an SVM and pattern matching study for detecting ADEs in Japanese discharge summaries [96].

Monitoring for Hospital-Acquired Infections
Healthcare-associated infections are a severe problem worldwide, and NLP methods show great potential to help in hospital surveillance [97], and predictions are proposed to be important in future proactive decision support for risk patients. The use of NLP in automated surveillance is partly driven by regulations [98-99] that in many countries require hospitals to report on adverse events, and several systems have been implemented with good results [100-103].

Discussion and Conclusion
In this survey, we outlined recent advances in clinical NLP for a multitude of languages with a focus on semantic analysis. Substantial progress has been made for key NLP sub-tasks that enable such analysis (i.e. methods for more efficient corpus construction and de-identification). Furthermore, research on (deeper) semantic aspects – linguistic levels, named entity recognition and contextual analysis, coreference resolution, and temporal modeling – has gained increased interest.

Current State of Clinical Semantic Analysis
Specifically, we have observed a great synergy within the research community to address the barriers of progress in clinical NLP addressed by both Meystræ et al. [4] and Chapman et al. [5], by developing and disseminating new textual and technical resources – in several languages, particularly through the creation of community shared tasks and adaptation of general NLP resources. Many NLP systems meet or are close to human agreement on a variety of complex semantic tasks. The clinical NLP community is actively benchmarking new approaches and applications using these shared corpora. In real-world clinical use cases, rich semantic and temporal modeling may prove useful for generating patient timelines and medical record visualizations, but may not always be worth the computational runtime and complexity to support knowledge discovery efforts from a large-scale clinical repository. For some real-world clinical use cases on higher-level tasks such as medical diagnosing and medication error detection, deep semantic analysis is not always necessary – instead, statistical language models based on word frequency information have proven successful. There still remains a gap between the development of complex NLP resources and the utility of these tools and applications in clinical settings.

Future Opportunities For Clinical NLP
Although there has been great progress in the development of new, shareable and richly-annotated resources leading to state-of-the-art performance in developed NLP tools, there is still room for further improvements. Resources are still scarce in relation to potential use cases, and further studies on approaches for cross-institutional (and cross-language) performance are needed. Furthermore, with evolving health care policy, continuing adoption of social media sites, and increasing availability of alternative therapies, there are new opportunities for clinical NLP to impact the world both inside and outside healthcare institution walls.

Bridging the Healthcare Policy and Practice Gap
There are new governmental policies and initiatives, e.g., Meaningful Use, advocating for next generation EHR technologies to en-
hance the efficiency and accuracy of healthcare delivery to the patient through clinical decision support, patient engagement, self-reported/ self-monitored data integration, and quality measure reporting [104]. For example, the Precision Medicine Initiative advocates for the development of tools to integrate patient genetic, environmental, and lifestyle data (e.g., data from medical/personal devices or social media) into the electronic medical record to support precision medicine e.g., patient-centered prevention, diagnostic, and treatment models for disease [105]. Similarly, the European Commission emphasizes the importance of eHealth innovations for improved healthcare in its Action Plan [106]. Such initiatives are of great relevance to the clinical NLP community and could be a catalyst for bridging health care policy and practice.

In conclusion, we eagerly anticipate the introduction and evaluation of state-of-the-art NLP tools more prominently in existing and new real-world clinical use cases in the near future.

Acknowledgements
We wish to thank Pierre Zweigenbaum, Aurélie Névélô, Wendy Chapman, and Stephane Meystre for valuable comments on this survey. This work was partially funded by Swedish Research Council (350-2012-6658), the Department of Veteran Affairs CREATE (CRE 12-312), National Library of Medicine (NLM R01LM010964) and the project High-Performance Data Mining for Drug Effect Detection at Stockholm University, funded by the Swedish Foundation for Strategic Research under grant IIS11-0053.

Integrating New Media for Accessing Population Health Status
Furthermore, with growing internet and social media use, social networking sites such as Facebook and Twitter have become a new medium for individuals to report their health status among family and friends. These sites provide an unprecedented opportunity to monitor population-level health and well-being, e.g., detecting infectious disease outbreaks, monitoring depressive mood and suicide in high-risk populations, etc. Additionally, blog data is becoming an important tool for helping patients and their families cope and understand life-changing illness.

Addressing Evolving Consumer Needs
Finally, with the rise of the internet and of online marketing of non-traditional therapies, patients are looking to cheaper, alternative methods to more traditional medical therapies for disease management. Little is understood about these interventions. NLP can help identify benefits to patients, interactions of these therapies with other medical treatments, and potential unknown effects when using non-traditional therapies for disease treatment and management e.g., herbal medicines.

References
26. Fernández O, South BR, Shen S, Meystre SM. A


47. MIMIC II: https://mimic.ph [16 June 2015]


p. 29-37.
89. ICD: http://www.who.int/classifications/icd/en/ [16 June 2015]

Correspondence to: Sumithra Velupillai Department of Computer and Systems Sciences Stockholm University Postbox 7003 164 07 Kista Sweden Tel: +46 8 161 174 Fax: +46 8 703 9025 E-mail: sumithra@dsv.su.se

Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis

IMIA Yearbook of Medical Informatics 2015