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

S. Velupillai[1], D. Mowery[2], B. R. South[2], M. Kvist[1,3], H. Dalianis[1]
[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, close 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

Yearb Med Inform 2015;10:183-93
http://dx.doi.org/10.15265/IY-2015-009
Published online August 13, 2015

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.

Background – Identifying Existing Barriers and Recent Developments that Support Semantic Analysis

In the comprehensive clinical NLP review by Meystre et al. in 2008 [4], a number of future challenges were mentioned, such as the need for more efficient methods for corpus creation in terms of effort required as well as patient confidentiality, the need for further research in complex semantic tasks such as discourse and temporality, and the need for improvements in system performance to enhance the acceptance of NLP in clinical research contexts. Similarly, the recent position paper by Chapman et al. in 2011 [5] enumerates barriers to clinical NLP progress, such as the lack of annotated training and benchmarking datasets, the lack of inexpensive and reliable de-identification techniques, and insufficient common clinical standards, and calls for more creative solutions to address these barriers. Additionally, the lack of resources developed for languages other than English has been a limitation in clinical NLP progress.

In recent years, the clinical NLP community has made considerable efforts to overcome these barriers by releasing and sharing resources, e.g., de-identified clinical corpora, annotation guidelines, and NLP tools, in a multitude of languages [6]. Moreover, the progress and amount of research on the essential steps enabling semantic analysis (i.e. corpus creation) has led to an increased number of studies on semantics, both on linguistic levels such as morphology and syntax, and on deeper semantic levels such as complex concept classification, co-reference resolution, and temporal reasoning. The development and maturity of NLP systems has also led to advancements in the employment of NLP methods in clinical research contexts.

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-

tions would be a considerable contribution to the development of semantic resources.

Pre-annotation, providing machine-generated annotations based on e.g. dictionary lookup from knowledge bases such as the Unified Medical Language System (UMLS) Metathesaurus [11], can assist the manual efforts required from annotators. A study by Lingren et al. [12] combined dictionaries with regular expressions to pre-annotate clinical named entities from clinical texts and trial announcements for annotator review. They observed improved reference standard quality, and time saving, ranging from 14% to 21% per entity while maintaining high annotator agreement (93-95%). In another machine-assisted annotation study, a machine learning system, RapTAT, provided interactive pre-annotations for quality of heart failure treatment [13]. This approach minimized manual workload with significant improvements in inter-annotator agreement and F1 (89% F1 for assisted annotation compared to 85%). In contrast, a study by South et al. [14] applied cue-based dictionaries coupled with predictions from a de-identification system, BoB (Best-of-Breed), to pre-annotate protected health information (PHI) from synthetic clinical texts for annotator review. They found that annotators produce higher recall in less time when annotating without pre-annotation (from 66-92%).

Another strategy to mitigate time and cost is to use different annotation methods for creating corpora meant to train a statistical system versus creating a gold standard. A recent study found that double annotation and consensus annotation are not always necessary when creating corpora to train a statistical system [15]. They showed that there was no statistically significant difference in results when training a model on single annotated data compared to double annotated plus consensus. A three step process is suggested: 1) use double annotation for only a small targeted sample to ensure guideline adherence, 2) allow annotators to work independently on different sections of the corpus, 3) train a machine learning model on the human annotations and apply this to a new dataset.

Scalability

Additionally, scalability can impose constraints upon the degree of semantic analysis. 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.

De-identification – Enabling Data Access and Modeling Semantic Entities

A consistent barrier to progress in clinical NLP is data access, primarily restricted by privacy concerns. De-identification methods are employed to ensure an individual's anonymity, most commonly by removing, replacing, or masking Protected Health Information (PHI) in clinical text, such as names and geographical locations. Once a document collection is de-identified, it can be more easily distributed for research purposes. Since the thorough review of state-of-the-art in automated de-identification methods from 2010 by Meystre et al. [21], research in this area has continued to be very active. The United States Health Insurance Portability and Accountability Act (HIPAA) [22] definition for PHI is often adopted for de-identification – also for non-English clinical data. For instance, in Korea, recent law enactments have been implemented to prevent the unauthorized use of medical information – but without specifying what constitutes PHI, in which case the HIPAA definitions have been proven useful [23].

Following the pivotal release of the 2006 de-identification schema and corpus by Uzuner et al. [24], a more-granular schema, an annotation guideline, and a reference standard for the heterogeneous MTSamples.com corpus of clinical texts were released [14]. The schema extends the 2006 schema with instructions for annotating fine-grained PHI classes (e.g., relative names), pseudo-PHI instances or clinical eponyms (e.g., Addison's disease) as well as co-reference relations between PHI names (e.g., John Doe *COREFERS* to Mr. Doe). The reference standard is annotated for these pseudo-PHI entities and relations. To date, few other efforts have been made to develop and release

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 → Arg1[patient/SBJ NP (experiencer)], Rel[undergo.1/VP (experience, undergo)], Arg2[chemotherapy/OBJ NP (experienced)]

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

[43]. 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, Dligach 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, experiencer, and temporality were developed through translations from English along with enhanced rules and regular expressions [55]. Final results for negation and experiencer were high (> 87% and > 99% F1, respectively), but lower for historical and hypothetical temporality properties (26-54% and 13-44% F1, respectively).

Experiencer 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 (Naive Bayes, k-Nearest Neighbours and Random Forest). There were no statistically significant differences in results for classifying experiencer 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

challenge [63-64]. For example, in contrast to previous corpora, the THYME events are encoded to the linguistic head of a phrase, and a new temporal expression type was introduced to capture time expressions such as *preoperative*. The THYME corpus was released as part of the SemEval-2015 Task 6: Clinical TempEval challenge [65], where the system approach results met or were close to human agreement on all subtasks except temporal relationships using machine learning approaches [66].

The first step in a temporal reasoning system is to detect expressions that denote specific times of different types, such as dates and durations. A lexicon- and regular-expression based system (TTK/GUTIME [67]) developed for general NLP was adapted for the clinical domain. The adapted system, MedTTK, outperformed TTK on clinical notes (86% vs 15% recall, 85% vs 27% precision), and is released to the research community [68]. In the 2012 i2b2 challenge on temporal relations, successful system approaches varied depending on the subtask. For instance, hybrid approaches combining rule-based systems such as HeidelTime [69], SUTIME [70] and GUTIME [67] with CRF or SVM machine learning models proved useful for time expression classification (up to 90% span F1), CRF models for event span detection (up to 92% F1), SVM models for event attribute detection (86% accuracy), while temporal relationships were classified with a variety of approaches, resulting in up to 68% F1 [63].

Other studies define coarser time representations. For instance, Raghavan et al. [71] created a model to distinguish time-bins based on the relative temporal distance of a medical event from an admission date (*way before admission*, *before admission*, *on admission*, *after admission*, *after discharge*). The model was evaluated on a corpus of a variety of note types from Methicillin-Resistant S. Aureus (MRSA) cases, resulting in 89% precision and 79% recall using CRF and gold standard features. In a study to classify patient history episodes in Bulgarian discharge notes, the authors defined rules to identify temporal markers (absolute or relative moments of time), resulting in 87% precision and 68% recall, and the direction of time for the episode starting point (backwards or forward) resulting in 74% precision [72].

Most studies on temporal relation classification focus on relations within one document. Cross-narrative temporal event ordering was addressed in a recent study with promising results by employing a finite state transducer approach [73].

Document-level Analysis

Other NLP annotation efforts aim to demonstrate the potential clinical utility of underlying semantic information for document-level analysis. Sentiment is well-studied in the general NLP domain, but not yet in the clinical domain. One note-worthy effort was the creation and release of guidelines and transcribed suicide notes to support emotion classification at the snippet (clauses and phrases) and document levels [74-75]. The dataset was released as part of the Fifth i2b2/VA/Cincinnati shared task. Another effort targeting clinical information retrieval problems was the 2011-2012 TREC medical records track [76] in which one of the largest datasets was distributed to the clinical NLP community through shared tasks [77].

To enable cross-lingual semantic analysis of clinical documentation, a first important step is to understand differences and similarities between clinical texts from different countries, written in different languages. Wu et al. [78], perform a qualitative and statistical comparison of discharge summaries from China and three different US-institutions. Chinese discharge summaries contained a slightly larger discussion of problems, but fewer treatment entities than the American notes. Social history was never documented in this corpus of Chinese notes.

A further level of semantic analysis is text summarization, where, in the clinical setting, information about a patient is gathered to produce a coherent summary of her clinical status. This is a challenging NLP problem that involves removing redundant information, correctly handling time information, accounting for missing data, and other complex issues. Pivovarov and Elhadad present a thorough review of recent advances in this area [79].

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 Symptomology

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

system [86]) improved determining mood states for diagnosing major depressive disorders compared to using diagnostic codes alone (area under receiver operating characteristic curve of 85-88% vs 54-55%) [87]. The underlying NLP methods were mostly based on term mapping, but also included negation handling and context to filter out incorrect matches.

NLP has also been used for mining clinical documentation for cancer-related studies. Spacíc et al. [88] present a review of current state-of-the art in this area, where they conclude that named entity recognition methods perform well (F1 between 80-90%) but that there is room for improvement for handling non-standard wordings and that the main bottleneck for progress in this area is the lack of available corpora.

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 subtasks 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 Meystre 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.

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.

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 Névéol, 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.

References

1. Zeng QT, Redd D, Divita G, Jarad S, Brandt C. Characterizing Clinical Text and Sublanguage: A Case Study of the VA Clinical Notes. J Health Med Inform 2011;S3:001
2. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 2002;35(4), 222-35.
3. Temnikova I, Nikolova I, Baumgartner Jr WA, Angelova G, Cohen KB. Closure Properties of Bulgarian Clinical Text. In Proc RANLP 2013 September:667-75.
4. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128-44.
5. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011 Sep-Oct;18(5):540-3.
6. Névéol A, H. Dalianis G. Savova, Zweigenbaum P. Didactic Panel: Clinical Natural Language Processing in Languages Other Than English. In: Proc AMIA Annu Symp 2014.
7. Zweigenbaum P, Névéol A; Section Editors for the IMIA Yearbook section on clinical natural language processing. Clinical NLP synopsis. Yearb Med Inform 2014; to appear
8. Cohen KB, Demner-Fushman D. Biomedical natural language processing (Vol. 11). John Benjamins Publishing Company; 2014.
9. Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005 May-Jun; 12(3): 296–8.
10. Pustejovsky J, Stubbs A. Natural Language Annotation for Machine Learning - A Guide to Corpus-Building for Applications. O'Reilly Media; 2012.
11. UMLS Metathesaurus: http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/ [16 June 2015]
12. Lingren T, Deleger L, Molnar K, Zhai H, Meinzen-Derr J, Kaiser Met al. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. J Am Med Inform Assoc 2014 May-Jun;21(3):406-13.
13. Gobbel GT, Garvin J, Reeves R, Cronin RM, Heavirland J, Williams J, et al. Assisted annotation of medical free text using RapTAT. J Am Med Inform Assoc 2014 Sep-Oct;21(5):833-41.
14. South BR, Mowery D, Suo Y, Leng J, Ferrández Ó, Meystre SM, et al. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. J Biomed Inform 2014 Aug;50:162-72.
15. Grouin C, Lavergne T, Neveol A. Optimizing annotation efforts to build reliable annotated corpora for training statistical models. The 8th Linguistic Annotation Workshop, 2-14 August 2014. ACL; 2014. p. 54-8.
16. Zhai H, Lingren T, Deleger L, Li Q, Kaiser M, Stoutenborough L, et al. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. J Med Internet Res 2013 Apr 2;15(4):e73.
17. Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform 2015 May 1.
18. Uzuner O, Solti , Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010 Sep-Oct;17(5):514-8.
19. Juckett D. A method for determining the number of documents needed for a gold standard corpus. J Biomed Inform 2012 Jun;45(3):460-70.
20. Gundlapalli AV, Redd A, Carter M, Divita G, Shen S, Palmer M, et al. Validating a strategy for psychosocial phenotyping using a large corpus of clinical text. J Am Med Inform Assoc 2013 Dec;20(e2):e355-64.
21. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010;10(1):70.
22. HIPAA: http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/index.html [14 December 2014]
23. Shin SY, Lyu Y, Shin Y, Choi HJ, Park J, Kim WS, et al. Lessons Learned from Development of De-identification System for Biomedical Research in a Korean Tertiary Hospital. Healthc Inform Res 2013 Jun;19(2):102-9.
24. Uzuner, Ö, Luo Y, Szolovits, Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007 Sep-Oct; 14(5): 550–63.
25. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc 2013; 20: 77–83.
26. Ferrández O, South BR, Shen S, Meystre SM. A

hybrid stepwise approach for de-identifying person names in clinical documents. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing; 2012 June. ACL, 2012. p. 65-72.
27. Meystre S, Ferrández O, Friedlin J, South BR, Shen S, Samore MH. Text de-identification for privacy protection : A study of its impact on clinical text information content. J Biomed Inform 2014;50:140-50.
28. Li M, Carrell D, Aberdeen J, Hirschman L, Malin BA. De-identification of clinical narratives through writing complexity measures. Int J Med Inform 2014 Oct;83(10):750-67.
29. Grouin C, Névéol A. De-identification of clinical notes in French: towards a protocol for reference corpus development. J Biomed Inform 2014 Aug;50:151-61.
30. Dalianis H, Velupillai S. De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields. J Biomed Semantics 2010 Apr 12;1(1):6.
31. Alfalahi A, Brissman S, Dalianis H. Pseudonymisation of person names and other PHIs in an annotated clinical Swedish corpus. In: Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012, May 26, Istanbul. P. 49-54.
32. Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2013 Jan 1;20(1):84-94.
33. Sánchez D, Batet M, Viejo A. Utility-preserving privacy protection of textual healthcare documents. J Biomed Inform 2014 Dec;52:189-98.
34. Fan JW, Yang EW, Jiang M, Prasad R, Loomis RM, Zisook DS, et al. Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences. J Am Med Inform Assoc 2013 Nov-Dec;20(6):1168-77.
35. Penn Treebank II guidelines: ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/root.ps.gz [16 June 2015]
36. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010 Sep-Oct;17(5):507-13.
37. Ferraro JP, Daumé H 3rd, Duvall SL, Chapman WW, Harkema H, Haug PJ. Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. J Am Med Inform Assoc 2013 Sep-Oct;20(5):931-9.
38. Laippala V, Viljanen T, Airola A, Kanerva J, Salanterä S, Salakoski T, et al. Statistical parsing of varieties of clinical Finnish. Artif Intell Med 2014 Jul;61(3):131-6.
39. Bretschneider C, Zillner S, Hammon M. Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach. In: Proceedings of BioNLP 2013. ACL; 2013. p. 27–35.
40. Albright D, Lanfranchi A, Fredriksen A, Styler WF 4th, Warner C, Hwang JD, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. J Am Med Inform Assoc 2013 Sep-Oct;20(5):922-30.
41. Wu ST, Kaggal VC, Dligach D, Masanz JJ, Chen P, Becker L, et al. A common type system for clinical natural language processing. J Biomed Semantics 2013 Jan 3;4(1):1.
42. CEM: http://wiki.siframework.org/file/view/CEReference20081114.pdf [16 June 2015]
43. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011 Sep-Oct;18(5):552-6.
44. Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, et al. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. J Am Med Inform Assoc 2014 Aug 21.
45. Elhadad N, Pradhan S, Lipsky Gorman S, Manandhar S, Chapman W, et al. SemEval-2015 Task 14: Analysis of Clinical Text. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). ACL; 2015.
46. Skeppstedt M, Kvist M, Nilsson GH, Dalianis H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. J Biomed Inform 2014 Jun;49:148-58.
47. MIMIC II: https://mimic.physionet.org/database.html [16 June 2015]
48. Mowery D, Velupillai S, South BR, Christensen L, Martinez D, Kelly L, et al. Task 2: ShARe/CLEF eHealth Evaluation Lab 2014. In: CEUR Workshop Proceedings 2014;1180.
49. Dligach D, Bethard S, Becker L, Miller T, Savova GK. Discovering body site and severity modifiers in clinical texts. J Am Med Inform Assoc 2014 May-Jun;21(3):448-54.
50. Xu J, Zhang Y, Wang J, Wu Y, Jiang M, Soysal E, et al. UTH-CCB: The Participation of the SemEval 2015 Challenge – Task 14. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). ACL; 2015.
51. Chapman B, Lee S, Kang H, Chapman W. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform 2011;44:728–37.
52. Velupillai S, Skeppstedt M, Kvist M, Mowery D, Chapman BE, Dalianis H, Chapman WW. Cue-based assertion classification for Swedish clinical text-developing a lexicon for pyConTextSwe. Artif Intell Med 2014 Jul;61(3):137-44.
53. Chapman WW, Hillert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, et al. Extending the NegEx lexicon for multiple languages. Stud Health Technol Inform 2013;192:677-81.
54. Mowery DL, Velupillai S, Chapman WW. Medical diagnosis lost in translation: analysis of uncertainty and negation expressions in English and Swedish clinical texts. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing; 2012 June. ACL; 2012. p. 56-64.
55. Afzal Z, Pons E, Kang N, Sturkenboom MC, Schuemie MJ, Kors JA. ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. BMC Bioinformatics 2014 Nov 29;15(1):373.
56. Cogley J, Stokes N, Carthy J, Dunnion J. Analyzing patient records to establish if and when a patient suffered from a medical condition. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing; 2012 June. ACL; 2012. p. 38-46.
57. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc 2012 Sep-Oct;19(5):786-91.
58. Bodnari A, Szolovits P, Uzuner Ö. MCORES: a system for noun phrase coreference resolution for clinical records. J Am Med Inform Assoc 2012 Sep-Oct;19(5):906-12.
59. Chowdhury MF, Zweigenbaum P. A controlled greedy supervised approach for co-reference resolution on clinical text. J Biomed Inform 2013 Jun;46(3):506-15.
60. Jindal P, Roth D. Using domain knowledge and domain-inspired discourse model for coreference resolution for clinical narratives. J Am Med Inform Assoc 2013 Mar-Apr; 20(2): 356–62.
61. Pustejovsky J, Lee K, Bunt H, Romary L. ISO-TimeML: An International Standard for Semantic Annotation. In: Proc 7th Intl Conference on Language Resources and Evaluation (LREC'10); 2012 May, Valletta, Malta. ELRA.
62. Styler WF, Bethard S, Finan S, Palmer M, Pradhan S, de Groen P, et al. Temporal annotation in the clinical domain. Transactions of the Association for Computational Linguistics. 2014;2:143–54.
63. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013 Sep-Oct;20(5):806-13.
64. Sun W, Rumshisky A, Uzuner O. Temporal reasoning over clinical text: the state of the art. J Am Med Inform Assoc 2013 Sep-Oct;20(5):814-9.
65. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. SemEval-2015 Task 6: Clinical TempEval. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). ACL.
66. Velupillai S, Mowery DL, Abdelrahman S, Christensen L, Chapman W. BluLab: Temporal Information Extraction for the 2015 Clinical TempEval Challenge. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). ACL.
67. Verhagen M, Mani I, Sauri R, et al. Automating temporal annotation with TARSQI. Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions. ACL; 2005. p. 81-4.
68. Reeves RM, Ong FR, Matheny ME, Denny JC, Aronsky D, Gobbel GT, et al. Detecting temporal expressions in medical narratives. Int J Med Inform 2013 Feb;82(2):118-27.
69. Strötgen J, Gertz M. Multilingual and Cross-domain Temporal Tagging. Language Resources and Evaluation 2013;47(2):269–98.
70. Chang AX, Manning CD. SUTIME: a library for recognizing and normalizing time expressions. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) 2012. ELRA; 2012.
71. Raghavan P, Fosler-Lussier E, Lai AM. Temporal classification of medical events. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing 2012 June. ACL; 2012.

p. 29-37.
72. Boytcheva S, Angelova G, Nikolova I. Automatic Analysis of Patient History Episodes in Bulgarian Hospital Discharge Letters. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23 - 27 2012. ACL; 2012. p. 77-81.
73. Raghavan, P, Fossler-Lussier E, Elhadad N, Lai A. M. Cross-narrative temporal ordering of medical events. In: Proc. ACL 2014. ACL; 2014. p. 998-1008.
74. Pestian JP, Matykiewicz P, Linn-Gust M. What's In a Note: Construction of a Suicide Note Corpus. Biomed Inform Insights 2012;5:1-6.
75. Pestian JP, Matykiewicz P, Linn-Gust M, South B, Uzuner O, Wiebe J, et al. Sentiment Analysis of Suicide Notes: A Shared Task. Biomed Inform Insights 2012; 5(Suppl 1): 3–16.
76. TREC Medical track: http://www-nlpir.nist.gov/projects/trecmed/2011/tm2011.html [16 June 2015]
77. Voorhees EM, Hersh W. Overview of the TREC 2012 Medical Records Track. In: Proc. TREC 2012.
78. Wu Y, Lei J, Wei WQ, Tang B, Denny JC, Rosenbloom ST, et al. Analyzing differences between Chinese and English clinical text: a cross-institution comparison of discharge summaries in two languages. Stud Health Technol Inform 2013;192:662-6.
79. Pivovarov R, Elhadad N. Automated Methods for the Summarization of Electronic Health Records. J Am Med Inform Assoc 2015.
80. Matykiewicz P, Cohen KB, Holland KD, Glauser TA, Standridge SM, Verspoor KM, Pestian J. Earlier Identification of Epilepsy Surgery Candidates Using Natural Language Processing. In: Proceedings of the 2013 Workshop on Biomedical Natural Language Processing. ACL; 2013.
81. Leeper NJ, Bauer-Mehren A, Iyer SV, LePendu P, Olson C, Shah NH. Practice-Based Evidence: Profiling the Safety of Cilostazol by Text-Mining of Clinical Notes. PLoS ONE 2013;8(5):e63499.
82. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 jun;13(6):395-405.
83. Shivade C, Raghavan P, Fosler-Lussier E, Elhadad N, Johnson SB, Lai AM. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21(2), 221–30.
84. Wright A, McCoy AB, Henkin S, Kale A, Sittig DF. Use of a support vector machine for categorizing free-text notes: assessment of accuracy across two institutions. J Am Med Inform Assoc 2013 Sep-Oct;20(5):887-90.
85. Yetisgen-Yildiz M, Bejan CA, Wurfel MM. Identification of Patients with Acute Lung Injury from Free-Text Chest X-Ray Reports. In: Proceedings of the 2013 Workshop on Biomedical Natural Language Processing. ACL; 2013.
86. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006 Jul 26;6:30.
87. Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med 2012 Jan;42(1):41-50.
88. Spasić I, Livsey J, Keane JA, Nenadić G, Text mining of cancer-related information: review of current status and future directions. Int J Med Inform 2014 Sep;83(9):605-23.
89. ICD: http://www.who.int/classifications/icd/en/ [16 June 2015]
90. Socialstyrelsen 2006. Socialstyrelsen - The National Board of Health and Welfare, Diagnosgranskningar utförda i Sverige 1997-2005 samt råd inför granskning, [Diagnostic reviews performed in Sweden 1997-2005 and advice for reviews] (In Swedish). http://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/9740/2006-131-30_200613131.pdf [12 June 2015]
91. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. A systematic literature review of automated clinical coding and classification systems. J Am Med Inform Assoc 2010;17(6):646-51.
92. Perotte A, Pivovarov R, Natarajan K, Weiskopf N, Wood F, Elhadad N. Diagnosis code assignment: models and evaluation metrics. J Am Med Inform Assoc 2014;21(2):231-7.
93. Martinez D, Cavedon L, Alam Z, Bain C, Verspoor K. Text mining for lung cancer cases over large patient admission data. In: Proceedings of the Abstracts of the Scientific Stream at Big Data 2014. CEUR Proceedings. p. 24-5.
94. Li Q, Melton K, Lingren T, Kirkendall ES, Hall E, Zhai H, et al. Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care. J Am Med Inform Assoc 2014 Sep-Oct;21(5):776-84.
95. Eriksson R, Bjødstrup Jensen P, Frankild S, Juhl Jensen L, Brunak S. Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. J Am Med Inform Assoc 2013; 20(5):947-53.
96. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki Ket al. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform 2010;160(Pt 1):739-43.
97. Freeman R, Moore LSP, García Álvarez L, Charlett A, Holmes A. Advances in electronic surveillance for healthcare-associated infections in the 21st Century: a systematic review. J Hosp Infect 2013 84(2):106-19.
98. EU HAI Guidelines: http://www.ecdc.europa.eu/en/activities/surveillance/HAI/Pages/default.aspx [16 June 2015]
99. USA HAI Guidelines: http://www.health.gov/hai/prevent_hai.asp [16 June 2015]
100. Shepard J, Hadhazy E, Frederick J, Nicol S, Gade P, Cardon A, et al. Using electronic medical records to increase the efficiency of catheter-associated urinary tract infection surveillance for National Health and Safety Network reporting. Am J of Infection Control 2014;42:e33-e36
101. Ehrentraut C, Kvist M, Sparrelid M, Dalianis H. Detecting Healthcare-Associated Infections in Electronic Health Records - Evaluation of Machine Learning and Preprocessing Techniques, in the Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM 2014); 2014, Aveiro, Portugal.
102. Proux D, Hagège C, Gicquel Q, Pereira S, Darmoni S, Segond F, et al. Architecture and systems for monitoring hospital acquired infections inside a hospital information workflow. In Proceedings of the Workshop on Biomedical Natural Language Processing; 2011 September, USA: Portland, Oregon.
103. Adlassnig K-P, Blacky A, Koller W. Artificial-Intelligence-Based Hospital- Acquired Infection Control. Strategy for the Future of Health. Stud Health Technol Inform 2009;149:103–10.
104. Meaningful use: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/Downloads/eQuality_EHR_TEP_Findings.pdf [16 June 2015]
105. The White House Office of the Press Secretary. FACT SHEET: President Obama's Precision Medicine Initiative. 2015 Jan 30. https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative [25 June 2015]
106. eHealth Action Plan 2012-2020 - Innovative healthcare for the 21st century. COM/2012/0736 final. http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52012DC0736&-from=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