# Clinical Natural Language Processing in 2015: Leveraging the Variety of Texts of Clinical Interest

A. Névéol, P. Zweigenbaum, Section Editors for the IMIA Yearbook Section on Natural Language Processing
LIMSI CNRS UPR 3251, Université Paris Saclay, Orsay, France

## Summary

**Objective**: To summarize recent research and present a selection of the best papers published in 2015 in the field of clinical Natural Language Processing (NLP).

**Method**: A systematic review of the literature was performed by the two section editors of the IMIA Yearbook NLP section by searching bibliographic databases with a focus on NLP efforts applied to clinical texts or aimed at a clinical outcome. Section editors first selected a shortlist of candidate best papers that were then peer-reviewed by independent external reviewers.

**Results**: The clinical NLP best paper selection shows that clinical NLP is making use of a variety of texts of clinical interest to contribute to the analysis of clinical information and the building of a body of clinical knowledge. The full review process highlighted five papers analyzing patient-authored texts or seeking to connect and aggregate multiple sources of information. They provide a contribution to the development of methods, resources, applications, and sometimes a combination of these aspects.

**Conclusions**: The field of clinical NLP continues to thrive through the contributions of both NLP researchers and healthcare professionals interested in applying NLP techniques to impact clinical practice. Foundational progress in the field makes it possible to leverage a larger variety of texts of clinical interest for healthcare purposes.

## Keywords

Awards and prizes; decision making, computer-assisted; medical informatics/trends; natural language processing; semantics

Yearb Med Inform 2016:234-9
http://dx.doi.org/10.15265/IY-2016-049
Published online November 10, 2016

## Introduction

The Clinical Natural Language Processing (NLP) section introduced in the 2015 IMIA Yearbook was continued in 2016. As in 2015, we define Clinical NLP as Natural Language Processing applied to clinical texts or aimed at a clinical outcome. This clearly encompasses NLP applied to texts in Electronic Health Records (EHRs), which is the case of the bulk of information extraction for decision support or clinical research. We also considered as clinically relevant applications research addressing the analysis of patient-authored text or speech for public health or diagnosis purposes. This year's survey paper of the NLP section reports on unintended consequences of clinical NLP [1]. It follows up with the observation we made in 2015 that there are two categories of efforts in clinical NLP: foundational methods and applications, with only budding impact on clinical practice. The best papers selected this year offer a good representation of the diversity of texts of clinical interest that clinical NLP addresses. Three papers seek to connect and aggregate multiple sources of information for a variety of clinical applications: unstructured and structured EHR data to detect medication discrepancies [2], expert-authored encyclopedic texts and EHR texts for document clustering [3], clinical trial descriptions and EHR texts for clinical trial eligibility [4]. Two papers study texts of clinical interest produced by patients, either in forums where they report experience on treatment [5], or in exams where texts elicited in controlled conditions can help diagnose mental disorders [6]. The best papers provide a contribution to methods [3], resources [5], applications [2], and sometimes a combination of these aspects [4, 6], which shows that the field is growing in maturity and breadth.

## About the Selection Process

Our method for pre-selecting papers was amended to increase the number of papers reviewed and avoid bias from the completion status of MEDLINE indexing. We relied on a three step process.

First, queries were sent to PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) and to the ACL Anthology Searchbench (http://aclasb.dfki.de/). The PubMed query used minimal metadata and free-text keywords: (English[LA] AND journal article[PT] AND 2015[dp] AND hasabstract[text]) AND ((medical OR clinical OR natural) AND "language processing"). The ACL Anthology query restricted our selection to the most selective conferences (ACL, EMNLP, NAACL, EACL, COLING, CONLL) and workshops (ACL BioNLP). It used the free text keywords medical, clinical and health. The systematic collection of papers through queries brought back 397 titles and abstracts from MEDLINE and 19 from ACL Anthology, resulting in a total selection of 416 papers, a 270% increase over the initial selection we made in 2015.

Second, we examined the retrieved titles and abstracts to select articles which met the following selection criteria: (1) NLP was performed on clinically relevant texts, (2) the work aimed at a clinical outcome, and (3) the contribution was significant from both the NLP and clinical standpoints. Review papers and correspondence were discarded

**Table 1** Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2016 in the section 'Natural Language Processing'. The articles are listed in alphabetical order of the first author's surname.

| Section<br>Natural Language Processing |
|---|
| ▪ Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, Tang H, Solti I, Ni Y. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC Med Inform Decis Mak 2015 May 6;15:37. |
| ▪ Luo Y, Xin Y, Hochberg E, Joshi R, Uzuner O, Szolovits P. Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. J Am Med Inform Assoc 2015 Sep;22(5):1009-19. |
| ▪ Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, Li Q, Zhai H, Solti I. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc 2015 Jan;22(1):166-78. |
| ▪ Oronoz M, Gojenola K, Pérez A, de Ilarraza AD, Casillas A. On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. J Biomed Inform 2015 Aug;56:318-32. |
| ▪ Prud'hommeaux E, Roark B. Graph-Based Word Alignment for Clinical Language Evaluation. Computational Linguistics 2015 Dec:41(4):549-78. |

at this stage (N=21) in order to keep only original research contributions. A significant number of papers (N=191) were also discarded because they were more focused on biological aspects (N=44), knowledge representation (N=14), or other topics such as psycholinguistics (N=83).

Third, selected articles were grouped by broad topics and ranked. We then browsed the full text of the top 24 papers to refine this pre-selection in order to cover each of the topics while ensuring that the final selection comprised a variety of topics, authors, and venues. In the list of references provided at the end of the synopsis, a star indicates papers that were in the final selection of 15 candidate best papers.

Articles were rated independently by the two section editors who reviewed the titles and abstracts and classified them in four categories: out of topic (OT), should be considered for best paper short list (Y), may be considered for best paper short list (M) and not a best paper (N). Inter-rater agreement on this classification was 0.71 per Cohen's kappa. Articles that received at least one Y or M label were discussed by both section editors to arrive at a consensual final selection.

It was decided to discard papers from the special issues related to NLP challenges (the JBI special issue covering i2b2 2014 [7], a JAMIA paper building on the ShARe/CLEF 2013 lab results [8]) and the temporal challenges results [9] because even though challenges provide valuable contributions to the field [10], challenge papers are usually polished working notes that report on work that has not reached the level of maturity expected from a "best paper". However, we do note that one i2b2 2014 track overview paper addressed a very interesting issue, namely assessing the ease of re-using clinical NLP systems [11]. This shows that the community is developing an awareness and interest to the need of building on each other's ground work.

## Topics in Clinical NLP

Based on the review of the papers published in 2015, we observed as in 2014 two clusters of publications: papers with a focus on NLP methods which can be applied to clinical information processing, and papers with a focus on applications of NLP of clinical interest.

## NLP Methods with Applications to Clinical Information Processing

The first cluster of publications focuses on natural language processing methods of interest to clinical text processing. While these methods encompass the multiple levels and dimensions of natural language processing, we observe renewed interest for processing speech input, continued efforts on negation detection, syntactic parsing, semantic role labeling, word sense disambiguation, and temporal analysis, and innovations in feature construction and selection for information extraction and text classification.

Suominen et al. [12] address information extraction from spoken nursing handover notes. They build a corpus of 100 documents and test several speech recognition and information extraction conditions. The task is challenging for automatic speech recognition and for evaluation because in addition to the usual difficulty of speech recognition, the spoken documents elaborate on the provided text documents, therefore there is not always a direct correspondence between text and speech.

We also mention below Xiao et al. [13] and Prud'hommeaux and Roark [6] who process health care professional speech and patient speech, respectively.

Negation detection is a common component in clinical information extraction (this is the case, e.g., in [5, 14, 15, 16]). Mehrabi et al. [17] specifically address negation detection with a focus on complex sentences. The method they propose augments NegEx with dependency parsing. It is initially trained and evaluated on a specialized domain where it successfully improves negation detection performance. However, when applied to a general clinical corpus from another hospital, it still improves precision but at the expense of recall. This confirms the conclusions of Wu et al. (2014) that "negation is not solved" [18].

Jiang et al. [19] compare the performance of three state-of-the-art parsers with and without specific training on a clinical treebank. They find that all parsers perform better with mixed training using a combination of Penn Treebank and clinical treebank. Wang et al. [20] perform additional tailored domain adaptation with the selective use of the Specialist lexicon and manual changes to the Stanford parser parameters based on corpus characteristics. However neither study shows the impact of the parsing improvement on a more applied task such as entity recognition.

Zhang et al. [21] compare the performance of three state-of-the-art semantic role labeling (SRL) algorithms on clinical texts of the MiPACQ corpus with and without domain adaptation using three out-of-domain corpora (PropBank, NomBank, and BioProp). They observe that the "feature

augmentation" domain adaptation method consistently brings a statistically significant improvement in F-measure. The best results are obtained when all available corpora are used for training. The study does not show the impact of the SRL improvement on a more applied task such as relation extraction.

Moon et al. [22] perform a pilot word-sense disambiguation study comparing three classifiers on seven clinical acronyms. The importance of this topic led Wu et al. [23] to build a prototype system which performs real-time clinical abbreviation recognition and disambiguation at the time clinical documents are written. This exemplifies how progress in fundamental methods can impact very concrete day-to-day clinical tasks.

Time is an important dimension in medical diagnosis and care and therefore in the analysis of patient data. Sun et al. [24] focus on the detection and normalization of relative and incomplete temporal expressions, which they estimate to 26% of temporal expressions based on the i2b2 2012 corpus. Lin et al. [25] address a phenotype whose definition comprises a temporal aspect: liver toxicity secondary to recent methotrexate treatment. Therefore, they take into account temporal features in their classifier. McCoy et al. [26], further described below, study patient re-admission, an event which might take place in the future. In contrast, most EHR classification tasks predict events or states which are current or past events, documented in the EHR documents to classify.

When applying machine learning to process language data, creating relevant features remains a complex task which requires a varying mix of domain knowledge, machine-learning know-how, time-consuming exploration of input data, and heuristic experimentation. Methods which help design better features are thus welcome. Meng and Morioka [27] automatically identify lexical patterns in clinical texts with alignment methods commonly used in computational biology and score them according to their positive predictive value for information extraction targets, such as entity types or relations between entities. The patterns may contain gaps which aim to help generalization, and generally perform better than two baseline methods to extract information from a corpus of chest radiology reports. Yu et al. [16] add to the input clinical texts a variety of online knowledge sources related to rheumatoid arthritis and coronary artery disease that are used as a proxy for experts. They extract UMLS concepts from these knowledge sources and select as features those which are neither too rare nor too common in the clinical texts and which exceed a correlation threshold with the target phenotype. Automatic features perform as well as expert-determined features when applied to identify cases of rheumatoid arthritis and coronary artery disease. Luo et al. [3], described in more detail below, elicit complex features based on frequent subgraphs by mining syntactic graphs whose nodes are UMLS concepts. These features are used in a novel unsupervised approach to clustering with potential application to EHR phenotyping. A remarkable property of these three methods is that they save system designers feature creation time and yet produce features which are humanly interpretable.

## Applications of NLP of Clinical Interest

The second cluster of publications focuses on applications of state-of-the-art NLP methods to clinical problems, which may involve non-NLP specialists confronted with practical clinical situations that can benefit from a contribution of NLP.

The detection of adverse drug reactions (ADR) in texts keeps momentum. Source data include patient forums [14, 15], tweets [15], and EHR texts [5]. Some systems detect actual relations beyond co-occurrence of drugs and disorders, possibly taking into account negations and knowledge about known drug indications, sometimes crossing sentence boundaries [5]. Definitions differ about which pieces of information are considered to detect an ADR relation. This can be an explicit mention of a relation between a drug and an effect [5, 14], or the mere mention of an effect in the context of a drug, where context is set for example by a thread about a drug in a forum, or a tweet with a drug hashtag [15]. The cited papers produce resources and tools to develop and test methods associated to these representations.

The detection of more complete information on drugs from heterogeneous sources, including dosage, drug form, administration route, frequency, strength, is covered in Li et al. [2]. ADR detection is also addressed in a more focused context by Lin et al. [9] for methotrexate-induced liver toxicity, mentioned below.

The computation of clinical trial eligibility is another application which remains active. Ni et al. [4] try to classify a complete patient record with respect to its eligibility for a clinical trial, based on information retrieval methods applied to word and concept vector representations of EHRs and eligibility criteria. Complementary to the previous approach, a more focused approach is provided by Shivade et al. [28] who identify sentences in EHRs that are relevant for making eligibility decisions.

Coding and classification of EHR texts has long been a target for natural language processing. Koopman et al. [29] revisit ICD10 coding on a previously unexplored type of documents viz. death certificates. They focus on cancer-related death and present a two-layer classification method which identifies whether or not a certificate is related to cancer and then classifies it according to cancer type. McCoy et al. [30] compute the similarity between patient records and Research Domain Criteria (RDoC) representative corpora; they show that this combination of information retrieval and NLP method can help transition from DSM-5 to RDoC coding. This topic will also be featured in the upcoming 2016 CEGS N-GRID shared task which targets the determination of symptom severity in an RDoC domain for a patient based on the text of their initial psychiatric evaluation.

EHR classification belongs to a "general trend of methodology development for phenotyping using the EMR data, including its free text" [9]. Joffe et al. [31] determine whether a patient diagnosis is positive or negative for breast cancer. They compare one-class vs two-class SVM classifiers and conclude that one-class is better for imbalanced datasets. Lin et al. [9] address the phenotyping of patients with a particular

ADR, methotrexate-induced liver toxicity in patients with rheumatoid arthritis. One originality is that their classifier takes into account temporal features. While the time of an event relative to the document date (DocTimeRel) adds 0.8pt of F-measure, the feature that contributes the most to the performance is section parsing (17pt). Xiao et al. [13] use Automatic Speech Recognition transcripts of clinician speech to rate the empathy of therapists during a therapy session with possible applications towards the training and evaluation of health care professionals. McCoy et al. [26] use psychiatric and general internal medicine discharge summaries to assign polarity scores for a positive/negative sentiment. A subsequent statistical analysis reveals that greater positive sentiment scores are associated with a reduction in readmission hazard in both cohorts. Prud'hommeaux and Roark [6] compute diagnostic scores from patient speech transcripts to help with dementia phenotyping; this work is described in greater detail below.

## Concluding Remarks

Languages other than English were less represented in 2015 than in 2014 in the papers we examined. As in 2014, the largest clusters in the application-oriented papers found when perusing the 2015 abstracts pertain to EHR classification (N=36) or information extraction from the EHR (N=29).

Interestingly, we observe a generalization of a vocabulary shift that has been emerging over the past few years from "EHR classification" to "phenotyping". In our opinion this reflects technical advances in genetic analysis and the increasing use of genotyping in clinical practice, which led to the adoption of the mirror term "phenotyping" when describing clinical observations.

Authors notably worked with a diversity of text types in addition to clinical texts: patient-authored texts continue to be an important source; transcribed audio was more present than in 2014; and online encyclopedic resources, such as Wikipedia or Medscape, are suggested to collect complementary knowledge to help analyze clinical texts. Research results show that each type of text has something to contribute to help analyze clinical information and build clinical knowledge.

## References

In the reference list below, papers that were shortlisted as best paper candidates are marked with a *.

1. Demner-Fushman D, Elhadad N. Aspiring to unintended consequences of Natural Language Processing: A Review of Recent Developments in Clinical and Consumer-Generated Text Processing. Yearb Med Inform 2016.
2. * Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC Med Inform Decis Mak 2015 May 6;15:37.
3. * Luo Y, Xin Y, Hochberg E, Joshi R, Uzuner O, Szolovits P. Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. J Am Med Inform Assoc 2015 Sep;22(5):1009-19.
4. * Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc 2015 Jan;22(1):166-78.
5. * Oronoz M, Gojenola K, Pérez A, de Ilarraza AD, Casillas A. On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. J Biomed Inform 2015 Aug;56:318-32.
6. * Prud'hommeaux E, Roark B. Graph-Based Word Alignment for Clinical Language Evaluation. Computational Linguistics 2015 Dec:41(4):549-78.
7. Journal of Biomedical Informatics. Volume 58, Supplement, December 2015, Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.
8. 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 2015 Jan;22(1):143-54.
9. Lin C, Dligach D, Miller TA, Bethard S, Savova GK. Multilayered temporal modeling for the clinical domain. J Am Med Inform Assoc 2016 Mar;23(2):387-95.
10. 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
11. Zheng K, Vydiswaran VG, Liu Y, Wang Y, Stubbs A, Uzuner Ö, et al. Ease of adoption of clinical natural language processing software: An evaluation of five systems. J Biomed Inform 2015 Dec;58 Suppl:S189-96.
12. * Suominen H, Zhou L, Hanlen L, Ferraro G. Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. JMIR Med Inform 2015 Apr 27;3(2):e19.
13. Xiao B, Imel ZE, Georgiou PG, Atkins DC, Narayanan SS. "Rate My Therapist": Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing. PLoS One 2015 Dec 2;10(12):e0143055.
14. Liu X, Chen H. A research framework for pharmacovigilance in health social media: Identification and evaluation of patient adverse drug event reports. J Biomed Inform 2015 Dec;58:268-79.
15. * Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015 May;22(3):671-81.
16. * Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. J Am Med Inform Assoc 2015 Sep;22(5):993-1000.
17. Mehrabi S, Krishnan A, Sohn S, Roch AM, Schmidt H, Kesterson J, et al. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. J Biomed Inform 2015 Apr;54:213-9.
18. Wu S, Miller T, Masanz J, Coarr M, Halgrim S, Carrell D, et al. Negation's not solved: generalizability versus optimizability in clinical natural language processing. PLoS One 2014 Nov 13;9(11):e112774.
19. Jiang M, Huang Y, Fan JW, Tang B, Denny J, Xu H. Parsing clinical text: how good are the state-of-the-art parsers? BMC Med Inform Decis Mak 2015;15 Suppl 1:S2.
20. * Wang Y, Pakhomov S, Ryan JO, Melton GB. Domain adaption of parsing for operative notes. J Biomed Inform 2015 Apr;54:1-9.
21. * Zhang Y, Tang B, Jiang M, Wang J, Xu H. Domain adaptation for semantic role labeling of clinical text. J Am Med Inform Assoc 2015 Sep;22(5):967-79.
22. Moon S, McInnes B, Melton GB. Challenges and practical approaches with word sense disambiguation of acronyms and abbreviations in the clinical domain. Healthc Inform Res 2015 Jan;21(1):35-42.
23. Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Song M, et al. A Preliminary Study of Clinical Abbreviation Disambiguation in Real Time. Appl Clin Inform 2015 Jun 3;6(2):364-74.
24. * Sun W, Rumshisky A, Uzuner O. Normalization of relative and incomplete temporal expressions in clinical narratives. J Am Med Inform Assoc 2015 Sep;22(5):1001-8.
25. * Lin C, Karlson EW, Dligach D, Ramirez MP, Miller TA, Mo H, et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. J Am Med Inform Assoc 2015 Apr;22(e1):e151-61.
26. * McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, Perlis RH. Sentiment Measured

in Hospital Discharge Notes Is Associated with Readmission and Mortality Risk: An Electronic Health Record Study. PLoS One 2015 Aug 24;10(8):e0136341.
27. Meng F, Morioka C. Automating the generation of lexical patterns for processing free text in clinical documents. J Am Med Inform Assoc 2015 Sep;22(5):980-6.
28. Shivade C, Hebert C, Lopetegui M, de Marneffe MC, Fosler-Lussier E, Lai AM. Textual inference for eligibility criteria resolution in clinical trials. J Biomed Inform 2015 Dec;58 Suppl:S211-8.
29. * Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. Automatic ICD-10 classification of cancers from free-text death certificates. Int J Med Inform 2015 Nov;84(11):956-65.
30. * McCoy TH, Castro VM, Rosenfield HR, Cagan A, Kohane IS, Perlis RH. A clinical perspective on the relevance of research domain criteria in electronic health records. Am J Psychiatry 2015 Apr;172(4):316-20.
31. Joffe E, Pettigrew EJ, Herskovic JR, Bearden CF, Bernstam EV. Expert guided natural language processing using one-class classification. J Am Med Inform Assoc 2015 Sep;22(5):962-6.

Correspondence to:
Aurélie Névéol, Pierre Zweigenbaum
LIMSI CNRS UPR 3251
Université Paris Saclay
Rue John von Neumann
91400 Orsay
France
E-mail: {neveol,pz}@limsi.fr

# Appendix: Content Summaries of Selected Best Papers for the 2016 IMIA Yearbook, section clinical Natural Language Processing

### Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, Tang H, Solti I, Ni Y

### An end-to-end hybrid algorithm for automated medication discrepancy detection

### BMC Med Inform Decis Mak 2015 May 6;15:37

This paper describes an operational approach to detect drug names and related information from electronic health records, in order to reconcile medication information in unstructured data (from the text of discharge summaries or discharge prescriptions) and structured data (from a health information database). This work is very clearly presented with helpful explanatory figures. It is targeted to a useful clinical application which has the potential for a significant impact. The authors use a state of the art machine learning system they developed in a prior work for extracting medication information and combine it with a rule-based reconciliation algorithm geared toward the application. While the methods used are not too complex, they are relevant and efficiently applied towards the clinical goal. The performance obtained is good as well as results in speeding up human monitoring. The authors also provide a good error analysis. Overall, this study is a very nice example of an application of state of the art NLP methods tailored to be used in clinical practice to achieve a clinical goal.

### Luo Y, Xin Y, Hochberg E, Joshi R, Uzuner O, Szolovits P

### Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text

### J Am Med Inform Assoc 2015 Sep;22(5):1009-19

This paper reports on the development of unsupervised complex statistical clustering methods for patient profiling. It models clinical narrative texts in the form of syntactic graphs whose nodes are UMLS concepts. These graphs are then mined to identify subgraphs and to compute their distribution in terms of context words and patients involved. The subgraphs are used to build clusters that can capture similar patients, similar sentence subgraphs, and similar words. The subgraphs are suitable for human perusal by medical professionals and can be interpreted to give clinical meaning to each of the clusters. This is very original work that is well described and well compared to the literature. The authors go beyond a formal evaluation of the method (which outperforms a series of strong baseline methods by at least 10 points of F-measure) and discuss the medical relevance of the discovered clusters. Although the experiments apply the proposed method to one type of clinical problem, lymphoma subtype identification, it shows the potential for generalization and for several concrete clinical applications, such as phenotyping, definition of phenotypes, perhaps even clinical trial eligibility. Overall, this paper is an excellent example of strong foundational clinical NLP method research with potential for future practical applications in a clinical or public health setting.

### Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, Li Q, Zhai H, Solti I

### Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department

### J Am Med Inform Assoc 2015 Jan;22(1):166-78

This paper presents state-of-the-art eligibility screening methods for the identification of patients to include in clinical trials. This is the first study to evaluate screening methods on large-scale real-world clinical data for randomly selected clinical trials. It measures impact on screening efficiency using a solid experimental protocol. The authors apply customized state of the art NLP methods and tools that they developed in a previous work. A specific pipeline is created to combine the tools adequately for this application. The approach relies both on structured data and clinical notes from electronic health records. The authors develop a gold standard real-world dataset and use it to perform a large-scale evaluation that yields impressive results with 450% increase in trial screening efficiency. Overall, this a clearly written report of a strong study addressing an important clinical problem with potentially high practical implications.

### Oronoz M, Gojenola K, Pérez A, de Ilarraza AD, Casillas A

### On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions

### J Biomed Inform 2015 Aug;56:318-32

This paper reports on the annotation of adverse drug reactions in clinical documents in Spanish. The study is conducted with a state of the art methodology: the corpus is pre-annotated for entities using a general analyzer adapted to the biomedical domain,

experts independently revise the pre-annotated entities (achieving high inter-annotator agreement) and add relation annotations. Finally a gold-standard is obtained through consensus reconciliation. A complex representation of adverse drug reactions is used in the form of explicit (drug, relation, adverse reaction) triplets. Additional advanced annotation features include marking negated, hypothetical, and discontinuous entities as well as inter-sentence relations. The authors successfully use the resource they created to train machine learning systems extracting entities and relations. This work also has to be commended for releasing the corpus under data agreement. Annotated clinical resources available to the community are still rare, especially for languages other than English.

### Prud'hommeaux E, Roark B

### Graph-Based Word Alignment for Clinical Language Evaluation

### Computational Linguistics 2015 Dec:41(4):549-78

This study describes the application of natural language processing approaches for studying phenomena associated with cognitive decline through the analysis of transcribed patient speech. Using text alignment techniques, the authors automatically compute a cognitive score that leads to diagnosis in combination with other clinical tests and exam results. The analysis carried out covers the structured content of the patient narrative instead of relying on global linguistic characteristics as was done previously. This clearly presented report introduces new methods and provides a comprehensive review of work addressing the analysis of patient narratives and pathologic language. It uses two existing datasets, including one that is publicly available. The authors show that their method is useful for computing an existing score defined in current protocol and also for deriving a new diagnosis tool. Overall, this is an innovative study with significant potential implications. The presentation and detail of the study is also quite exemplary of work leveraging informatics approaches in clinical contexts.