Automated Diagnostic Code Extraction in Electronic Medical Records

Nicholas J Hotz
Indiana University
nhotz@iu.edu

ABSTRACT

Electronic medical records (EMRs) play an increasingly important role in healthcare. However, the rapidly growing volume of text in EMRs creates challenges for information extraction (IE). As such, many research institutions are developing computer-based systems to automate EMR structured IE. This paper investigates the processes, the challenges, and the current state of automated IE of EMRs with a specific focus on automated systems that extract ICD9 codes from clinical text. While automated system performance has caught up to the accuracy of manual coding under specific circumstances, automated code extraction remains mostly an academic exercise. To extract value from their work, researchers should shift their focus away from highly specialized algorithms that work in isolation and instead collaborate with industry to develop augmented intelligence systems that help make coding professionals more effective.

KEYWORDS

Natural Language Processing, Medical NLP, Clinical NLP, Information Extraction, Clinical Coding, Healthcare Big Data

1 INTRODUCTION

Demand for structured health data continues to grow [20], and the adoption of electronic health records (EMRs) generates new opportunities to improve clinical care, administrative processes, clinical workflows, and patient outcomes through higher quality, more accurate, more consistent, and more easily accessible documentation [14] [17].

However, the size, growth, and textual nature of EMRs render traditional software and hardware unable to effectively manage healthcare big data [18]. Healthcare data in the United States reached 150 exabytes in 2011 with Kaiser Permanente, Califronia's health network, reportedly having between 26.5 and 44 petabytes alone [5]. The volume of healthcare data is doubling every 12-14 months [7], and the diversity of this data further complicates its analysis [10]. Much of it is stored in narrative form which describe patients, their own and their family's medical history, their personal lifestyle, and their current medical conditions [14]. Although convenient for documentation, narrative text is difficult for computer systems to interpret as coded data that can support research, provide clinical knowledge and performance information, and improve patient outcomes [14] [20].

Commonly studied clinical NLP problems include de-identification [23], the development of patient problem summaries [8], and diagnostic code extraction [15]. This paper focuses on diagnostic code extraction which is the process of converting EMR clinical narratives into appropriate medical codes such as ICD9 (the standard medical diagnostic hierarchical taxonomy system in the United States until September 30, 2015). Perotte et al. describe that both the ICD9 and the more recently adopted ICD10 taxonomies as ”organized in a rooted tree structure, with edges representing is-a relationships between parents and children” [15]. Kavauluru et al. explain that the ICD9 and ICD10 leaf nodes are codes that provide specific information used for ”billing and reimbursement, quality control, epidemiological studies, and cohort identification for clinical trials” [12].

Currently, coding professionals and clinicians manually extract diagnostic codes from EMRs which is expensive, inefficient, and has become increasingly complex due to various factors including the expansion of payment systems, new reporting requirements, increased oversight and regulation, and the increased volume of EMR data [1] [17] [20] [23]. This complexity limits manual coding accuracy. Manual coders often disagree [16] and are more specific than sensitive in their code assignments [3]. Errors are prevalent; for example, a Swedish study of 4,200 patient records found errors in 20% of the main diagnoses [23]. Over-coding can lead to fraud if healthcare providers bill for services not rendered while under-coding prevents providers from earning reimbursements for valid conditions and services [15].

Since the 1990s, researchers have tried to improve the coding processes through automated coding and classification technologies [11]. Stanfill et al. in their comprehensive literature review in 2010 describe these automated coding systems as ”a variety of computer-based approaches that transform narrative text in clinical records into structured text, which may include assignment of codes from standard terminologies, without human interaction”. They cite that the American Health Information Management Association asserted in 2004 that, ”The industry needs automated solutions to allow the coding process to become more productive, efficient, accurate, and consistent”. Yet, Stanfill et al. conclude that the relative performance of automated systems to manual coding is not yet known [20]. As of 2008 and still in 2015, automated systems are still mostly used for research purposes with few applications in use by practitioners [14] [23].

2 EMR INFORMATION EXTRACTION CHALLENGES

Several challenges have slowed the development of clinical text NLP applications, which lag behind NLP applications in other fields [4]. Meystre, et al attribute the lack of shareable clinical data as the biggest challenge [14]. Large annotated corpora are needed to develop effective machine learning algorithms that can classify roughly 17,000 possible ICD-9 codes and 68,000 ICD-10 codes whose frequency distributions are highly skewed [2]. However, clinical information needs to be de-identified (which itself is a challenging problem) in order to comply with privacy concerns and regulations such the USA's Health Information Portability and Protection Act (HIPAA) and the European Union's General Data Protection Regulation (GDPR); as a consequence, large corpora typically remain siloed within individual healthcare systems and are rarely available for outsiders [14] [20].

As a related problem, even when corpora are available, the annotation process is time-consuming, expensive, and traditionally relies on domain experts and linguists [14] [23]. Given the highly specific sublanguages of clinical text, general NLP systems perform poorly on cross-domain clinical texts without these comprehensive annotated corpora. Consequently, much of the development in clinical text NLP occur in siloes and is not used outside of the laboratory in which they were developed [4].

In addition to the lack of shared annotated corpora, Meystre et al. present four challenges that hinder the development of effective clinical text IE. First, clinical narratives contain ungrammatical phrases with short-hand abbreviations and acronyms. About a third of these short-hand texts are overloaded (a single unit may have multiple meanings) which can be challenging for human interpretation and even more challenging for computer interpretation. Second, the rate of misspellings is around 10% [19], which is higher than most texts is difficult for many NLP techniques. Third, clinical texts often contain long series of non-text information, such as laboratory test results, which makes sentence segmentation difficult. Forth, institution-specific pre-formatted templates that appear in clinical texts are difficult for interpretation and their meanings do not transfer to other institutions' information [14]. Chapman et al. discuss additional challenges including the inadequacy of de-identification algorithms, the lack of focus for NLP in non-English clinical texts, and the absence of common clinical standards [4].

Fortunately, recent progress is promising as explained in literature reviews by Delanis et al (2014) and Velupillai et al (2015). These publications praise the clinical NLP community for overcoming many of these hurdles by providing more annotated corpora, developing more advanced NLP tools specific to clinical text, leveraging partially-automated processes to facilitate the annotation of corpora, and focusing on multiple languages [6] [23].

3 EMR PRE-PROCESSING

To convert text to medical codes, clinical text flows through various pre-processing and context feature detection techniques. General pre-processing NLP tools are being adopted and specialized for medical texts including:

- **Language Detection:** Multi-lingual studies may start with language detection algorithms, although some might still rely on manual detection [8].
- **Spell checking:** Clinical NLP spell checking uses standard dictionaries and medical-specific tools such as unified medical language system (UMLS) and WordNet [14].
- **Word sense disambiguation:** WSD allows the system to identify the correct meaning of a word that has multiple definitions; however this process is not as accurate with clinical texts as with general English (about 90% for general English and 80% for clinical text) [14].
- **Tokenization and sentence-splitting:** Tokenization is the process for breaking text into tokens such as words, phrases, or symbols [8] [21].
- **Part-of-speech tagging:** Also known as lexical analysis, POS tagging identifies a word's part of speech and its relationship with other words in a sentence [8] [14].
- **Parsers:** Parsers identify the sentence syntax, word dependencies, and expressions of interest [8] [14].

Context feature detection and analysis happen concurrently or following the above steps and identify how words and concepts are used in the context of the sentence. Clinical NLP systems often use a set of regular expressions and algorithms such as NegEx, NegExpander, TimeText, and ConText to define feature context. Notable contexts are negation (e.g. patient *does not* have a condition), speculative (e.g. patient *might have* a condition) temporality (e.g. to identify if the patient *has* or *had* a condition), subject identification (e.g. to identify if the condition belongs to the patient or someone else such as a family member), and severity (such as mild, moderate, or severe conditions) [14] [23].

4 REVIEW OF AUTOMATED ICD9 CODE EXTRACTION EFFECTIVENESS

To evaluate the effectiveness of automated systems, studies compare evaluation metrics against standards. Per Stanfill et al.'s literature review of 113 studies, 43% of studies use the gold standard comparison which uses two or more independent reviewers with an adjudication process for disagreements, and 51% use the regular practice standard of one reviewer [20]. Although considered more reliable, gold standards are still prone to error [15]. The most commonly reported metrics include recall or sensitivity (69%), PPV or precision (46%), specificity (43%), and accuracy (25%) [20].

Most studies focus only on a specific subset of clinical texts or diagnoses such as subdomains like radiology [17], for specific diagnoses like congestive heart failure [9] or cancer [13], or to extract only attributes of patients like smoking status [22]. Although many of these studies achieve accuracy metrics comparable or even exceeding gold standards, their results are not generalizable for more comprehensive or practical purposes in the field [20].

However, two recent studies attempt to comprehensively extract ICD9 codes from large EMR sets. In 2013, Perotte et al. attempted to extract ICD9 codes from the clinical text of Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC II), a publicly available database containing de-identified records of 40,000 ICU hospital admissions. They split the 22,815 discharge summaries, which contain 215,826 ICD9 codes (5030 distinct) into 20,533 training documents and 2,282 testing documents. Using a hierarchy support vector machine (SVM) classifier, they achieved an F-measure of 39.5% with a 30.0% recall and 57.7% precision. They also attempted a flat SVM which returned a 27.6% F-measure with 16.4% recall but with a higher precision (86.7%) [15].

Similarly, in 2015 Kavuluru et al. developed automated coding systems with 71,463 in-patient EMRs from the University of Kentucky Medical Center. They conclude that the best-performing automated coding method depends on the size and characteristics of the dataset. For smaller narratives in subdomains such as radiology or pathology, chain classifiers perform best because codes are highly related to each other. However, feature and data selection methods perform best with more comprehensive in-patient EMRs. Meanwhile, ”for large EMR datasets, the binary relevance approach with learning-to-rank based code reranking offers the best performance”. They reported a micro F score of 0.48 with codes that occur at last 50 times and a score of 0.54 for codes that occur in at least 1% of records [12].

5 OUTLOOK

Researchers are increasingly studying clinical NLP and diagnostic code extraction. However, the output of most research is limited to specific circumstances and has not yet been applied to practical use cases that improve the accuracy and efficiency of medical coding processes. Rather, the research community seems to evaluate its work in terms of algorithm accuracy metrics in their specific strength zones relative to the performance of human coders. Cross-domain medical coding studies are a step in the right direction toward a more practical approach which begins to mimic the reality faced by human coders.

However, the clinical NLP researchers should take this progress further, and collaborate with software engineers, HCI design specialists, business analysts, medical coders, and clinicians to develop practical augmented intelligence systems. These systems, which can include semi-automated recommendation and auditing support software solutions, can aid medical coding professionals in actual workflows to extract diagnostic codes from medical text. A workflow that leverages the strengths of algorithmic systems to shore up areas of human coder weaknesses can optimize medical coding efficiency and accuracy.

REFERENCES

[1] 2013. Automated Coding Workflow and CAC Practice Guidance (2013 update). (11 2013). http://bok.ahima.org/PB/CACGuidance#.WchAZMiGOUl

[2] Stefan BERNDORFER and Aron Henriksson. 2017. Automated Diagnosis Coding with Combined Text Representations. *Informatics for Health: Connected Citizen-Led Wellness and Population Health* 235 (2017), 201.

[3] Elena Birman-Deych, Amy D Waterman, Yan Yan, David S Nilasena, Martha J Radford, and Brian F Gage. 2005. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical care* 43, 5 (2005), 480–485.

[4] Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D’avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. (2011).

[5] Mike Cottle, Waco Hoover, Shadaab Kanwal, Marty Kohn, Trevor Strome, and N Treister. 2013. Transforming Health Care Through Big Data Strategies for leveraging big data in the health care industry. *Institute for Health Technology Transformation, http://ihealthtran. com/big-data-in-healthcare* (2013).

[6] Hercules Dalianis, Aurélie Névéol, Guergana Savova, and Pierre Zweigenbaum. 2014. Didactic Panel: clinical Natural Language Processing in Languages Other Than English. In *AMIA Annual Symposium 2014*. American Medical Informatics Association, S–84.

[7] Ivo D Dinov. 2016. Volume and value of big healthcare data. *Journal of medical statistics and informatics* 4 (2016).

[8] Crescenzo Diomaiuta, Maria Mercorella, Mario Ciampi, and Giuseppe De Pietro. 2017. A novel system for the automatic extraction of a patient problem summary. In *Computers and Communications (ISCC), 2017 IEEE Symposium on*. IEEE, 182–186.

[9] Jeff Friedlin and Clement J McDonald. 2006. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. In *AMIA annual symposium proceedings*, Vol. 2006. American Medical Informatics Association, 269.

[10] Sullivan Frost. 2015. Drowning in big data? reducing information technology complexities and costs for healthcare organizations. (2015).

[11] Ramakanth Kavuluru, Sifei Han, and Daniel Harris. 2013. Unsupervised extraction of diagnosis codes from EMRs using knowledge-based and extractive text summarization techniques. In *Canadian Conference on Artificial Intelligence*. Springer, 77–88.

[12] Ramakanth Kavuluru, Anthony Rios, and Yuan Lu. 2015. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial intelligence in medicine* 65, 2 (2015), 155–166.

[13] Burke W Mamlin, Daniel T Heinze, and Clement J McDonald. 2003. Automated extraction and normalization of findings from cancer-related free-text radiology reports. In *AMIA Annual Symposium Proceedings*, Vol. 2003. American Medical Informatics Association, 420.

[14] Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, John F Hurdle, et al. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 35, 128 (2008), 44.

[15] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2013. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association* 21, 2 (2013), 231–237.

[16] John P Pestian, Christopher Brew, Pawe l Matykiewicz, Dj J Hovermale, Neil Johnson, K Bretonnel Cohen, and W lodzis law Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. Association for Computational Linguistics, 97–104.

[17] Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors. 2016. Natural language processing in radiology: a systematic review. *Radiology* 279, 2 (2016), 329–343.

[18] Wullianallur Raghupathi and Viju Raghupathi. 2014. Big data analytics in healthcare: promise and potential. *Health information science and systems* 2, 1 (2014), 3.

[19] Patrick Ruch, Robert Baud, and Antoine Geissbühler. 2003. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial intelligence in medicine* 29, 1 (2003), 169–184.

[20] Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association* 17, 6 (2010), 646–651.

[21] Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. Sentence and token splitting based on conditional random fields. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*. 49–57.

[22] Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association* 15, 1 (2008), 14–24.

[23] Sumithra Velupillai, D Mowery, Brett R South, Maria Kvist, and Hercules Dalianis. 2015. Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of medical informatics* 10, 1 (2015), 183.

6 BIBTEX ISSUES

Warning–entry type for ”AHIMA” isn’t style-file defined

–line 64 of file report.bib

Warning–no key, author in AHIMA

Warning–to sort, need author or key in AHIMA

Warning–no key, author in AHIMA

Warning–no key, author in AHIMA

Warning–no key, author in AHIMA

Warning–empty author in AHIMA

Warning–no number and no volume in cottle2013transforming

Warning–page numbers missing in both pages and numpages fields in cottle2013transforming

Warning–empty publisher in dalianis2014didactic

Warning–empty address in dalianis2014didactic

Warning–page numbers missing in both pages and numpages fields in dinov2016volume

Warning–empty publisher in diomaiuta2017novel

Warning–empty address in diomaiuta2017novel

Warning–empty publisher in friedlin2006natural

Warning–empty address in friedlin2006natural

Warning–empty publisher in kavuluru2013unsupervised

Warning–empty address in kavuluru2013unsupervised

Warning–empty publisher in mamlin2003automated

Warning–empty address in mamlin2003automated

Warning–empty publisher in pestian2007shared

Warning–empty address in pestian2007shared

Warning–empty publisher in tomanek2007sentence

Warning–empty address in tomanek2007sentence

(There were 24 warnings)

7 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

7.1 Formatting

HID and i523 not included in the keywords

7.2 Character Errors

Erroneous use of quotation marks, i.e. use “quotes” , instead of ” ”

7.3 Structural Issues

Acknowledgement section missing

change section title OUTLOOK to CONCLUSION