Original Article

Artificial Intelligence-Driven Structurization of Diagnostic Information in Free-Text Pathology Reports

Pericles S. Giannaris[1,2], Zainab Al-Taie[1,5], Mikhail Kovalenko[1,2], Nattapon Thanintorn[2], Olha Kholod[1,2], Yulia Innokenteva[1], Emily Coberly[2], Shellaine Frazier[2], Katsiarina Laziuk[2], Mihail Popescu[4,1,3], Chi-Ren Shyu[1,3], Dong Xu[3,1], Richard D. Hammer[2,1], Dmitriy Shin[2,1,3*]

[1]Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, [2]Department of Pathology and Anatomical Sciences, School of Medicine, University of Missouri, Columbia, MO 65212, [3]Department of Electrical Engineering and Computer Science, College of Engineering, University of Missouri, Columbia, MO 65211, [4]Department of Health Management and Informatics, School of Medicine, University of Missouri, Columbia, MO 65212, United States, [5]Department of Computer Science, College of Science for Women, University of Baghdad, Baghdad, Iraq

Received: 22 May 2019 Accepted: 18 December 2019 Published: 11 February 2020

Abstract

Background: Free-text sections of pathology reports contain the most important information from a diagnostic standpoint. However, this information is largely underutilized for computer-based analytics. The vast majority of NLP-based methods lack a capacity to accurately extract complex diagnostic entities and relationships among them as well as to provide an adequate knowledge representation for downstream data-mining applications. **Methods:** In this paper, we introduce a novel informatics pipeline that extends open information extraction (openIE) techniques with artificial intelligence (AI) based modeling to extract and transform complex diagnostic entities and relationships among them into Knowledge Graphs (KGs) of relational triples (RTs). **Results:** Evaluation studies have demonstrated that the pipeline's output significantly differs from a random process. The semantic similarity with original reports is high (Mean Weighted Overlap of 0.83). The *precision* and *recall* of extracted RTs based on experts' assessment were 0.925 and 0.841 respectively (P <0.0001). Inter-rater agreement was significant at 93.6% and inter-rated reliability was 81.8%. **Conclusion:** The results demonstrated important properties of the pipeline such as *high accuracy, minimality* and *adequate knowledge representation*. Therefore, we conclude that the pipeline can be used in various downstream data-mining applications to assist diagnostic medicine.

Keywords: Free-text pathology reports, information extraction, *n*-ary modeling, structurization

Introduction

Pathologists document the diagnostic process of complex cancer diagnosis in unstructured and semistructured free-text pathology reports. These documents provide diagnostic information including clinical history, immunophenotypes, complex morphological features, and various molecular and genomic tests such as fluorescent *in situ* hybridization, cytogenetics, and next-generation sequencing.[1,2]

A thorough analysis of diagnostic information is critical to make a correct diagnosis. For instance, the diagnosis of classical Hodgkin lymphoma (cHL) is based, to a great extent, on the combination of morphology and immunophenotypic biomarkers (e.g., CD20 and CD30).[2,3] To computationally analyze diagnostic information to improve the process of diagnosis, information is required to be in a structured format. The demand for structured representation of diagnostic data is exemplified by pathologists' increasing use of computerized college of american pathologists (CAP) checklists, synoptic reporting (summarizations), semistructured final diagnosis, current procedural terminology coding (CPT) systems, tumor node metastasis (TNM) cancer staging systems, systematized nomenclature of medicine-clinical terms (SNOMED), cancer registry data and validation systems, or patient stratification techniques.

However, the bulk of biomedically significant information is stored in free-text format without any predefined structure. This

Access this article online

Quick Response Code:	
	Website: www.jpathinformatics.org
	DOI: 10.4103/jpi.jpi_30_19

Address for correspondence: : Dr. Dmitriy Shin,
Department of Pathology and Anatomical Sciences, School of Medicine,
University of Missouri, Columbia, MO 65212, USA.
E-mail: shindm@health.missouri.edu

For reprints contact: reprints@medknow.com

How to cite this article: Giannaris PS, Al-Taie Z, Kovalenko M, Thanintorn N, Kholod O, Innokenteva Y, *et al*. Artificial intelligence-driven structurization of diagnostic information in free-text pathology reports. J Pathol Inform 2020;11:4.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2020/11/1/4/278125

makes computational analysis of data and their relationships across pathology reports a daunting task.[4] In addition to that, the current attempts to present diagnostic findings in structured format (e.g., TNM cancer staging, synoptic reporting, or checklists) are unable to fully express the biological complexity of involved diagnostic entities (DEs). To this end, it is important to extract information from unstructured free-text pathology reports. Structured representation of diagnostic data could facilitate the development of knowledge bases (KBs), knowledge graphs (KGs), summarization applications, as well as question and answering systems. Published research illustrates that structured data can enable applications that could potentially enhance services and research related to patient cohort identification, discovery of predictive biomarkers for precision medicine, the study of mechanisms underlying cancer genesis for treating individual patients, or cancer surveillance to name a few.[1,5-18]

Currently, methods from natural language processing (NLP) and information extraction (IE) fields are used to convert information to a computable form. These NLP-IE techniques analyze natural language text and attempt to output data in structured form (e.g., vector and subject–predicate–object triples). The contribution of NLP-IE for extracting complete sets of information from biomedical text is being mostly demonstrated by machine learning (automatic model building for data analysis) and rule-based (*if-then* algorithm statements) approaches.[7,19-28] However, in the vast majority of cases, the extraction of information is limited to relatively simple DEs, including various cancer characteristics such as tumor site, stage, and diagnosis. To increase the chance of extraction of all important information contained in free text (high *recall*), recent NLP research has been extended to introduce the open IE (openIE) paradigm. This is a self-supervised learning task, which aims at the extraction of all possible relations between data in a text.[29,30] As such, this approach has potential to be more suitable for the extraction of diagnostic information. In openIE methods, the relations between DEs are usually expressed as subject-predicate-object relational triples (RTs), for instance, [PAX5]-(shows bright positivity in)-[B-cells], or [heterogeneous cell population]-(are composed of)-[small to medium sized round lymphocytes].[30-36] Such triples can generate KGs.

However, the output of the state-of-the-art openIE approaches, often, include "*uninformative extractions*" (i.e., extractions that omit critical information), "*incoherent extractions*" (i.e., extractions where the relational phrase has no meaningful interpretation), and "*overly-specific relations that convey too much information to be useful in further downstream semantic tasks*".[30] For example, from the following example of a pathology report, "the large neoplastic cells show bi- and multi-nucleation with large nuclei, pale chromatin, prominent cherry red nucleoli, and abundant cytoplasm consistent with Reed-Sternberg cell variants [. . .]. Neoplastic cells are also negative for ALK1, EBV, CD57, EMA, and CD7", openIE applications would generate the following RTs: (large neoplastic cells)-(*show*)-(multi-nucleation with large nuclei, pale chromatin, prominent cherry red nucleoli), (Neoplastic cells)-(*are negative for*)-(ALK1, EBV, CD57, EMA), (cells)-(*show*)-(large). Note here that the above output contains triples that suffer from compoundness, i.e., presence of several DEs in either the subject and/or the object of an RT, which is also referred as a lack of minimality in the openIE literature.[30] Yet, another drawback is an incoherent extraction i.e., (large)-(*show*)-(neoplasticl). Therefore, such compound RTs makes it impossible to generate KGs and mine them for implicit disease patterns. To achieve such computational analyses, RTs should express atomic information,[30] for example, (Neoplastic cells)-(*are negative for*)-(ALK1) or (neoplastic cells)-(*show*)-(multinucleation with large nuclei).

Here, we introduce a novel informatics pipeline to extract information from free-text pathology reports as sets of atomic RTs. To accomplish that, we extend a state-of-the-art openIE method by adding two critical steps of (i) atomization of compound RTs and (ii) their knowledge representation using *n*-ary relational modeling. The next section provides a detailed description of the methodology.

Methods

Overview

Our structurization pipeline consists of two main processes: Foreground process (FP) and background process (BP) [Figure 1]. In a semiautomated and recurring BP, pathology reports are searched for simple DEs (e.g., *Reed-Sternberg cells*), phrases that represent complex DEs (e.g., *perivascular fibrosis with "onion skinning"*) as well as other terms needed for structurization (e.g., *Mayo Clinic-Rochester Main Campus*). These terms are then added to a diagnostic practice vocabulary (DPV) and linked to appropriate terms in a diagnostic practice ontology (DPO). In addition, reports are searched for relations that can serve as predicates in RTs.

Thereafter, in the first two steps [P1 and P2 in Figure 1] of a FP, a cohort of pathology reports for a specific study is retrieved and preprocessed. To extract information in RT format, we employ the state-of-the-art Stanford OpenIE[37] application, for which at scheduled times in the BP, the Stanford Named Entity Recognition[38] classifier is retrained to recognize named entities in pathology reports [P3, P6 in Figure 1]. In the final step of the FP, the subject and/or object of Stanford OpenIE-generated compound RTs are split into atomic terms, which are identified using DPV. Then, for each Stanford OpenIE-generated RT, a *n*-ary model, utilizing atomic terms, is created. Such *n*-ary model reflects the same semantics as the original RT. A set of *n*-ary models derived from a free-text pathology report represents a fully structurized version of a report.

Data acquisition and preprocessing (foreground process)

We have queried the pathology department's medical records' systems to generate a cohort of free-text pathology reports for our analysis (See section S2 in supplementary material [SM]). Since natural language in pathology reports is often

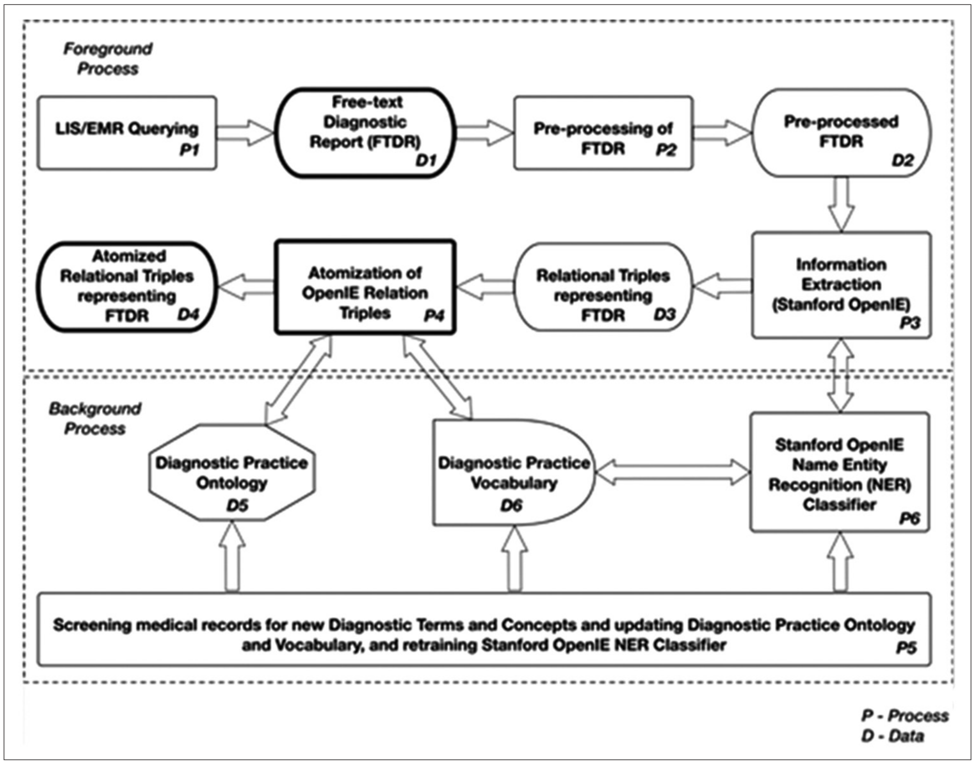

Foreground Process
LIS/EMR Querying P1
Free-text Diagnostic Report (FTDR) D1
Pre-processing of FTDR P2
Pre-processed FTDR D2
Information Extraction (Stanford OpenIE) P3
Relational Triples representing FTDR D3
Atomization of OpenIE Relation Triples P4
Atomized Relational Triples representing FTDR D4
Background Process
Diagnostic Practice Ontology D5
Diagnostic Practice Vocabulary D6
Stanford OpenIE Name Entity Recognition (NER) Classifier P6
Screening medical records for new Diagnostic Terms and Concepts and updating Diagnostic Practice Ontology and Vocabulary, and retraining Stanford OpenIE NER Classifier P5
P - Process
D - Data

Figure 1: Architecture of extraction and structurization of diagnostic information pipeline

characterized by long sentences (e.g., “Reed-Sternberg cells are negative for CD20, CD3, CD43, BCL6, CD79a, EMA, Alk-1, and LCA [CD45]”), phrases that use multiple verbs (e.g., “Flow cytometric analysis reveals B-cells show no evidence of a monotypic population or aberrant antigen expression”), incomplete expressions (e.g., “no clusters are present,” “no blasts”), variable punctuation and symbols (e.g., “\”, “;”, “:”, “+/-”), abbreviations (e.g., *RS-cells* instead of *Reed-Sternberg cells*), etc., it is challenging for openIE applications to efficiently extract information. To address this issue, we preprocess reports with a combination of regular expression scripts to (i) remove patient, physician, and document identifiers, (ii) split reports into sentences, (iii) standardize variations of proper names, etc., (e.g., “CJC, PSF” replaced by “chris j chris psf,” “Dr. Smith” replaced by “john smith md”), (iv) edit from variable to single space in text, (v) convert punctuation marks to “,” or “.”, and (vi) convert all letter character strings to lower case,[39-41] [S3 in Supplementary Material].

Diagnostic practice ontology and vocabulary (background process)

Our method utilizes the DPO, a schema to represent concepts from pathology practice, and a corresponding DPV, which includes instantiations of concepts from DPO. The DPV consists of various terms from pathology reports detected and identified by the BP. To refer to specific diseases and diagnostic tests, concepts from standard ontologies and controlled vocabularies such as SNOMED (e.g., ICD codes) and Cluster Differentiation System (e.g., names of antibody tests like CD30) are included in DPO. Names of biomedical providers and healthcare system departments, as well as descriptions such as “*onion skinning*,” “*soccer ball-like*,” “*popcorn cells*,” “*hyperlobate cells in a ‘shotgun distribution’*,” used by pathologists in a specific pathology department, are encoded in the DPV.

Concepts and terms in the DPO and the DPV are logically organized as follows. First of all, DPO consists of concepts that represent general terms in a pathology practice. For instance, persons and organizations are represented as a super class Healthcare_Actor along with corresponding sub-classes such as Pathologist, Cytologist, Health_Organization. Second of all, DPO contains classes corresponding to such concepts as specimen, biomarkers, which are represented by a super class Healthcare_Object and its specific sub-classes such as Specimen and Biomarker.

DPV and DPO are updated manually or with computer scripts as new diagnostic specialists are hired; new biomedical techniques are used, or new terms are found in reports. For example, for a newly hired pathology resident “*Dr. Smith, MD*,” an instance of a class Pathology_Resident will be added to the DPV. However, if the job status of “*Dr. Smith*” changes to Attending_Pathologist, the corresponding instance will be updated. The ontology was created through a reverse engineering process during which pathologists, technologists, and staff in a pathology practice were interviewed.

Named Entity Recognition and open information extraction technologies (background process)

The purpose of modules P3 and P6, Figure 1, is to extract RTs from free-text diagnostic reports using openIE technologies. First of all, to achieve this, the Stanford's NER classifier is employed [P6 in Figure 1] to label sequences of words in free-text diagnostic reports that are names of things and assist openIE.[42-44] The classifier is trained with an expanded vocabulary of named entities from 135 diagnosis comments on various cancer types and 35 microscopic descriptions on cHL. These documents contain a relatively large number of *named entities* such as names of medical providers, organizations, as well as names of various DEs. In our pipeline, sequences of words are labeled based on the following predefined categories: (i) person (e.g., [Charles J Chris PSF]$_{person}$, [Miranda D Crown TRANSCRIBER]$_{person}$), (ii) organization (e.g., [Department of Pathology]$_{organization}$, [University of Missouri Healthcare System]$_{organization}$), (iii) location (e.g., [1 Hospital Dr, Columbia, MO 65201, Suite N224]$_{location}$), and (iv) immunophenotypic (e.g., [germinal center b-cell phenotype]$_{immunophenotypic}$, [follicular dendritic cells]$_{immunophenotypic}$). Then, a set of 136 diagnostic reports is used to train the classifier. Another 34 diagnostic reports are used as a testing set to evaluate the predictive performance of the classifier on the labeled named entities.

Second, the Stanford OpenIE technology software is used to extract RTs. Specifically, this tool extracts all possible RTs discovered in a text without relying on predefined text patterns or a pre-specified relation schema.[29,43,45] The extracted RTs are in *subject-predicate-object* form as it is generally described in the literature.[36,45,46]

Structurization process (foreground process)

The structurization process starts with the detection, identification and atomization of compound Stanford OpenIE-generated RTs. For example, an RT extracted from example (a) in Table 1, is compound because the object represents multiple terms describing different properties of a cell population, namely, "*large folded multilobulated nucleus*," "*inconspicuous nucleolus*," "*scant cytoplasm*," "*popcorn cell variant*," "*lymphocyte-predominant-lp cell variant*." Similarly, in the RT extracted from example (b) in Table 1, the object represents two immunohistochemical (IHC) antibody tests: "cd30" and "cd15." The general workflow of the structurization process is depicted in Figure 2.

To detect compound triples and identify minimal tokens that represent DEs or named entities, the pipeline utilizes a word-alignment method with the Sliding Window technique to match words or phrases in RTs to terms in the DPV [See algorithm for Vocabulary Matching procedure in Figure 2 in S4 of Supplementary Material]. The Sliding Window scans words in the subject and/or object of a RT. Each time the window slides into a word the program checks whether the phrase in the window corresponds to a term in the DPV. The procedure continues until the Sliding Window has covered all words in the RT. In this manner, a set of tokens is generated. If both the subject and object are minimal, the RT is considered to be atomic and is added to the KG representation of the report. Otherwise, the RT is marked as compound and passed into the atomization and *n*-ary modeling procedure [See algorithm in Figure 3 in S4 of Supplementary Material]. Here, ontology patterns are leveraged in order to develop the *N*-ary Relation Modeling and to link an entity to multiple other entities.[47]

The atomization and *n*-ary modeling procedure takes a set of subject and object tokens generated by the Vocabulary

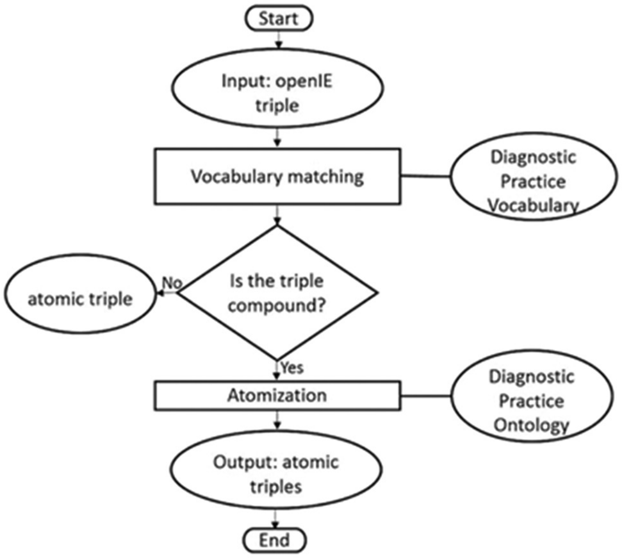


Figure 2: Workflow of the structurization process

Table 1: Examples of compound relational triples

(a) "...neoplastic cells have a single large folded multilobulated nucleus with an inconspicuous nucleolus and scant cytoplasm consistent with a popcorn or lymphocyte-predominant-lp cell variant. …"		
	Extracted relational triples	
Subject	**Predicate**	**Object**
neoplastic cells	have	large folded multilobulated nucleus with inconspicuous nucleolus consistent with lymphocyte-predominant-lp cell variant
	Extracted relational triples	
(b) "… neoplastic cells are positive for cd30 and cd15 …"		
Subject	**Predicate**	**Object**
neoplastic cells	are positive for	cd30 cd15

Matching and initiates a *N*-ary Relational Modeling. For this, an appropriate *N*-ary Anchor as well as a set of *N*-ary Predicates are retrieved from the DPO. The *N*-ary Predicates link each token of the subject and object to the *N*-ary Anchor. The *N*-ary Anchor is selected according to the relational predicate of the RT [line 4 in Figure 3 in S4 of Supplementary Material]. Each predicate is encoded as an ontological relation in the DPO. In Supplementary Material, the algorithm provides steps to select *N*-ary Predicates to link tokens to the *N*-ary Anchor [lines 7, 8 in Figure 2 in S3 of Supplementary Material]. *N*-ary Anchors and *N*-ary Predicates are encoded manually in the BP.

The set of all *n*-ary modeling representations of all RTs from a pathology report constitutes a KG of the report. We use a a Resource Description Framework (RDF) store for storage and retrieval of the integrated KG generated for a set of pathology reports of a specific study.

To illustrate the structurization process, consider an example of free-text pathology report:

Neoplastic cells are negative for CD45, CD20, BCL-6, CD10, CD23, and ALK. MUM-1 and CD79a also highlight plasma cells.

First of all, we preprocess the free-text report with our regular expression scripts (See S3 in SM) to convert all characters to lower case and to singular from, and to remove punctuation, and stop words. Thereafter, we segment the text into sentences to get the following text:

- neoplastic cell are negative for cd45 cd20 bcl6 cd10 cd23 alk
- mum1 cd79a highlight plasma cell.

Table 2: Relational triples extracted from the example in the in the case illustration

RT	Subject	Predicate	Object
1	neoplastic cell	are negative for	cd45 cd20 bcl-6 cd10 cd23 alk
2	mum-1 cd79a	highlight	plasma cells

RT: Relational triple

Next, the Stanford OpenIE is utilized to extract RTs, which are presented in Table 2.

In the third step, the subject and object of each triple are tokenized using the Vocabulary Matching procedure [Figure 2 in Supplementary Material], with each token having a matching term in the DPV. For instance, the following two sets of tokens are generated for the subject and object of the first RT:

- SUBJECT: *neoplastic cell*$_{(token\_1)}$
- OBJECT: *cd45*$_{(token\_1)}$, *cd20*$_{(token\_2)}$, *bcl-6*$_{(token\_3)}$, *cd10*$_{(token\_4)}$, *cd23*$_{(token\_5)}$, *alk*$_{(token\_6)}$.

We have to emphasize here that the DPV is produced in the BP. It contains terms that strictly represent named entities, such as *CD20*, *germinal center B-cell phenotype*. However, in some cases, there could be two DEs where one is an *extension* of the other. For instance, there could be two terms, *neoplastic cell* and *neoplastic cell proliferation*. The Vocabulary Matching procedure is biased towards finding the most *complete* term in the DPV to match adjacent words in a RT. To this end, the Sliding Window does not stop when it finds a matching term for neoplastic cell in the DPV but continues to process the next word. If it finds a term *neoplastic cell proliferation* in the DPV, it generates a token for this phrase. Otherwise, a token for *neoplastic cell* is generated.

In the next step, since both RTs are compound, with the first having a compound object and the second a compound subject, they are passed to the atomization and *n*-ary modeling procedure [Figure 3 in S5 of Supplementary Material]. The first step in this procedure is to determine a *n*-ary anchor for the atomization of the RT. For the relational predicates "are negative for" and "highlight", *N*-ary Anchor "IHC_Study" is retrieved from the DPO, where it was encoded before this step in the BP. During the second step, *N*-ary_Predicates "is_done_for" and "is_negative_for" are retrieved for the subject and the object of the first RT, and "is_positive_for" and "is_done_for" for the subject and the object of the second RT correspondingly. The resulting *n*-ary modeling

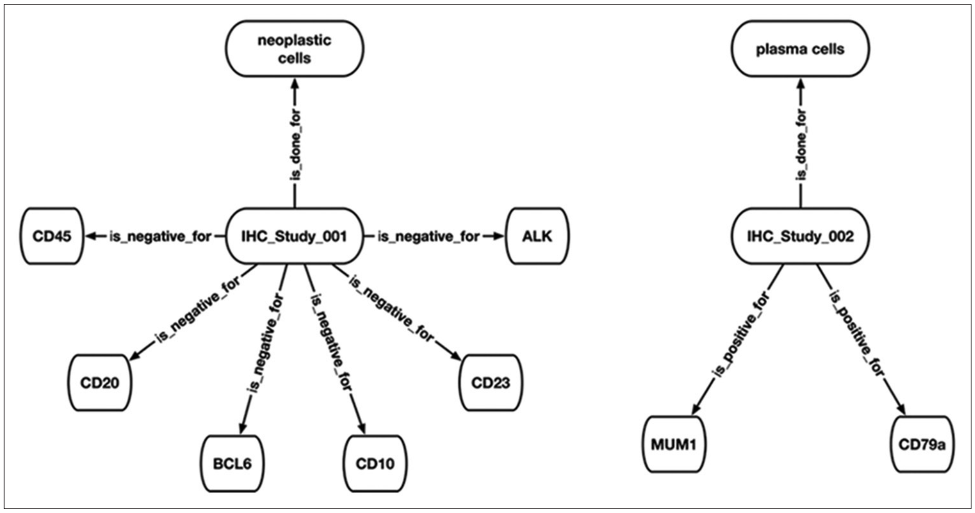


Figure 3: *N*-ary modeling representation of the free-text pathology report in the case illustration

representation of the original pathology report constitutes a KG [Figure 3].

Evaluation metrics

The effectiveness of the structurization pipeline is evaluated by the performance measures of *precision* and *recall*. Specifically, these measures correspond to the pipeline's capacity to generate all correct output (*recall*/sensitivity) "*and only the correct output (precision/*positive predictive value*)*".[43] These measures are adapted from the information retrieval and openIE fields.[48-50] In addition, comparisons with performances of other openIE applications extend this evaluation process.

Results and Discussion

Evaluation studies

We have conducted several studies to evaluate our structurization pipeline with a sample of 35 microscopic description pathology reports that describe cases of Hodgkin's Lymphoma from 2014 to 2017. These reports included bone marrow and lymph node biopsies and a variety of laboratory tests (e.g., immunohistochemistry, flow cytometry, and molecular genetics). The reports were written in a narrative style that describes specimen of adult patients of different age, sex, and race. For additional information, see S2 in SM. The next sections provide detailed descriptions and results of these studies along with a discussion of the properties and limitations of the pipeline.

Comparison with a random process

First, we compared RTs generated by our pipeline with a set of randomly generated RTs to determine that the pipeline follows certain behavior that is distinctive from a uniform behavior.

To conduct this evaluation, we randomly selected a data sample of 115 typical RTs produced by our algorithm from a set of 592 RTs. Then, we uniformly randomly sampled terms in DPV and predicates from DPO to construct a set of 136 *random RTs*. The random RTs were added to the 115 RTs from the data sample to construct a *combined set of 251 RTs*. Thereafter, a discrete probability mass function (PMF) was derived by dividing the number of instances of each RT by the size of the combined set of RTs. This PMF reflected data generation logic of the structurization algorithm. The reference PMF was generated as a uniform PMF for the *combined set of 251 RTs*.

We used the Kullback–Leibler Divergence (KLD) to measure the difference between the two distributions.[40] The KLD value was 1.236. *KLD* values approximating zero denote significant similarity between two distributions.[51,52] Since resulting KLD value was not close to zero, we concluded that the pipeline's output differed from the random process.

The two distributions have been also found statistically different with a two-sample Kolmogorov–Smirnov test *D*-statistic of 0.916 and p-value < 0.0001. *D*-statistic values below 0.1 identify matching distributions.[53] The results indicated that the RTs generated by the structurization process are statistically different from RTs generated by a uniform random process.[54-56]

Semantic similarity assessment

In the second study, we computed the semantic similarity between the original 35 free-text pathology reports and their corresponding KGs generated by the structurization pipeline. To achieve this, we utilized the *Align Disambiguate and Walk* (ADW) tool that computes the similarity between two lexical items such as words, sentences, and documents.[57,58] ADW is a knowledge-based system that leverages the Topic-Sensitive PageRank algorithm over a graph of word senses generated by the WordNet ontology.[59,60] We calculated the semantic similarity between the two datasets with the ADW adaptations of the *Cosine*, a measure of the angle between two documents represented as vectors with numerical values, as well as *Weighted Overlap* (WO), a similarity measure between "*diverse and dissimilar inputs in order to create an integrated analysis*", measures. We report the mean values of these computations. Specifically, mean *Cosine* similarity is at 0.77 with a standard deviation of 0.063. Here, as values approach 1 the smaller the angle thus, greater the similarity. Mean WO is at 0.83 with a standard deviation of 0.051. Here, as values approach 1 the greater the overlap between the inputs thus, greater the similarity. The Cosine method measures the semantic similarity between two documents at the *word level*. On the other hand, the WO method calculates the similarity between two documents by *considering larger text elements such as sentences, paragraphs or entire documents*. Since we analyzed biomedical information that depends on the context of an entire pathology report, we accepted the WO score as a measure of semantic similarity.

The mean WO score of 0.83 indicated that 17% of information was not found to be semantically similar. Further analysis of that information which was not included into the resulting KGs revealed that it consisted mostly of diagnostically irrelevant information such as case numbers and dates. After removal of such information from the original reports, we recalculated the WO similarity score. The mean WO score increased an average of 6%. Therefore, we deemed that the resulting KGs are semantically similar to the original free-text pathology reports in terms of diagnostic information. Note that although semantic similarity increased, 11% of information is not found to be semantically similar. This is attributed to invalid, uninformative, and incoherent RTs generated by the underlying openIE application. Consequently, our framework does not analyze that information.

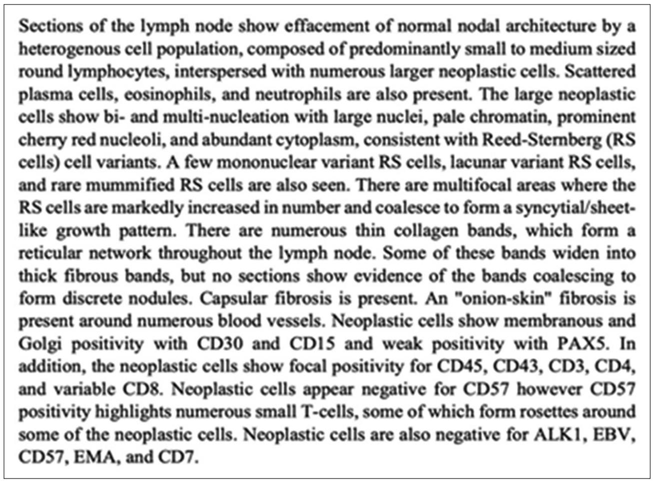
Performance measures

Comparison of current open information extraction applications

Our preliminary experiments with existing openIE methods showed that they cannot be directly used to extract diagnostic information in pathology domain. Table 3 shows comparative results of extraction of diagnostic information from free-text microscopic description sections of pathology reports. To compute *precision* and *recall*, we manually counted the number of distinct diagnostic informational points (DIPs) (e.g., "*CD20 is positive*") in the free-text reports [Figure 4] and the number of distinct DIPs expressed by the *subject-predicate-object* triples in the output of the openIE pipelines. As we expected, the openIE pipelines generated large numbers of RTs, with Stanford OpenIE and ClausIE being ahead of the other methods. However, the number of coherent RTs in the extractions was significantly lower than the overall number of extracted RTs, which resulted in low *recall* values, even in the best cases. It happened because of the *redundancy* resulted from the presence of *uninformative* and *incoherent* RTs, which was in turn caused by the pipelines' inability to accurately extract complex DEs (low *precision*). Even when DEs were correctly extracted, the resulting RTs suffered from *compoundness*, i.e., presence of several DEs in either the *subject* and/or the *object* of an RT, which is also referred as a lack of *minimality* in the openIE literature.[30] Yet, another drawback of existing openIE methods is unsatisfactory retention of the context in extracted RTs. Only some of the above-mentioned methods offer extraction of the contextual information and even in such cases the context often lacks adequate knowledge representation for effective and efficient implementation of downstream data-mining tasks (e.g., KGs). For an example of an extraction by Stanford OpenIE that illustrates the discussed shortcomings, (See S1 in SM).

Table 3: Comparative results of extraction of diagnostic information from free-text microscopic description section of pathology reports by different open information extraction methods

	Stanford OpenIE	OLLIE	ClauseIE	CSD	ReVerb
Precision	9.80%	18.18%	23.81%	14.29%	16.67%
Recall	18.52%	7.41%	18.52%	11.11%	3.70%



Sections of the lymph node show effacement of normal nodal architecture by a heterogenous cell population, composed of predominantly small to medium sized round lymphocytes, interspersed with numerous larger neoplastic cells. Scattered plasma cells, eosinophils, and neutrophils are also present. The large neoplastic cells show bi- and multi-nucleation with large nuclei, pale chromatin, prominent cherry red nucleoli, and abundant cytoplasm, consistent with Reed-Sternberg (RS cells) cell variants. A few mononuclear variant RS cells, lacunar variant RS cells, and rare mummified RS cells are also seen. There are multifocal areas where the RS cells are markedly increased in number and coalesce to form a syncytial/sheet-like growth pattern. There are numerous thin collagen bands, which form a reticular network throughout the lymph node. Some of these bands widen into thick fibrous bands, but no sections show evidence of the bands coalescing to form discrete nodules. Capsular fibrosis is present. An "onion-skin" fibrosis is present around numerous blood vessels. Neoplastic cells show membranous and Golgi positivity with CD30 and CD15 and weak positivity with PAX5. In addition, the neoplastic cells show focal positivity for CD45, CD43, CD3, CD4, and variable CD8. Neoplastic cells appear negative for CD57 however CD57 positivity highlights numerous small T-cells, some of which form rosettes around some of the neoplastic cells. Neoplastic cells are also negative for ALK1, EBV, CD57, EMA, and CD7.

Figure 4: Example of a pathology report demonstrating (i) complex diagnostic entities, (ii) complex relations among these diagnostic entities, and (iii) context in which these complex relations take place

Precision, recall, and inter-rater reliability

We assessed the effectiveness of the structurization pipeline with statistical measures of performance from information retrieval,[61] and openIE fields.[37,43,50] To do that, we recruited six domain experts (pathologists and bioinformaticians) from the University of Missouri Department of Pathology to evaluate RTs generated by the structurization procedure. All experts were familiar with the concept of the RTs and were asked to provide evaluations in a 3-point Likert-like scale:[62] Score 1 - *the triple does not state a fact from the pathology report*, Score 2 - *the triple "somewhat" states a fact from the pathology report*, and Score 3 - *the triple accurately states a fact from the pathology report*. Specifically, we measured *precision* and *recall* based on the number of *informational points* corresponding to a set of 3,836 RTs generated by the structurization pipeline and the number of *informational points* determined by a panel of diagnostic experts through analysis of the original pathology reports. Here, an *informational point* in a text is a single fact or a diagnostic finding, such as the presence or absence of a specific cell type or result of a test, reflected in a pathology report. Therefore, we define a RT as *relevant* to a diagnostic report if it corresponds to an *informational point* in the report. As such, *precision* is the ratio of *relevant* RTs that are returned by the pipeline over all RTs returned by the pipeline [Equation 1]. *Recall* is the ratio of relevant RTs that are returned by the pipeline over all *informational points* in the pathology report [Equation 2]. Here, a relevant RT is a RT that has received the top score 3 from all six domain experts. According to our experts, this type of RT correctly corresponded to an *informational point* in the corresponding pathology report. RTs evaluated with scores 2 or 1 were considered *ambiguous* and/or *incomprehensive* respectively. The total number of RTs refers to all the RTs evaluated with scores 3, 2, and 1. For *recall,* we counted all relevant RTs divided by a composite denominator, which equals to the sum of relevant RTs plus the number of *informational points in the reports that were not extracted by the pipeline as RTs* and marked as *"missed."*

$$Precison = \frac{RTs\ with\ score\ 3}{Sum\ of\ RTs\ with\ scores\ 1{,}2{,}3} \qquad \text{(Equation 1)}$$

$$Recall = \frac{RTs\ with\ score\ 3}{RTs\ with\ score\ 3 + \text{"}missed\text{"}\ RTs} \qquad \text{(Equation 2)}$$

We achieved *precision* of 0.925, which means that 92.5% of informational points of the pathology reports have been extracted by the structurization pipeline and expressed as RTs in the corresponding KGs. *Recall* was 0.841, which means that the pipeline has the capacity to generate relevant RTs. We used Fisher's exact test to test the hypothesis that the proportion of RTs with score 3 was statistically different than the proportion of RTs with score 1 and 2, for which the following contingency table was constructed [Table 4].

Fisher's exact test has odds ratio 19.7 and $P < 0.0001$. Therefore, we rejected the null hypothesis and concluded that the pipeline generates RTs that have high probability of

Table 4: Contingency table for the Fisher's exact test

RTs	RTs with score 3	RTs with score <3
Returned	TP: 3,551	FP: 69
Missed	FN: 485	TN: 186

TP: True positives, FN: False negatives, TN: True negatives, FP: False positives, RTs: Relational triples

being evaluated with the score 3. Our experiment specifically converts pathology reports rich in diagnostic information to machine-readable RTs. Since the RTs express stand-alone facts from the data, they are independent of each other. In such case, we are interested in the exact probability of whether RTs are associated or not regardless the sample size.[63-66]

To assess the agreement among experts in this study we computed inter-raters' reliability (IRR) score according to a *two-way random effects model based on a fully crossed design* as described in.[55] To do that, we computed the intra-correlation coefficient (ICC) statistic that reflects the level of correlation and magnitude of agreement between domain experts,[67-69] based on recommendations by McGraw and Wong.[70] Accordingly, we selected a single score intra-class correlation, absolute agreement, two-way random effects model with six raters across 3,836 triples, with average-measure ICC as our IRR measures. The ICC score was 0.818, with *P* value of 0.99. We, therefore, do not reject the *null hypothesis* and concluded that the differences in the assessment were *statistically insignificant*. According to Cicchetti's study, we consider ICC values between 0.80 and 0.90 as *good* and anything above as *excellent*.[67,71,72] Additionally, the calculated percentage of agreement was 93.6%. This statistic expresses the percentage of evaluations in which the domain experts are in absolute agreement.[69,71,73] ICC and percent agreement were computed in R using the "*irr*" package.[74] These results indicate high level of agreement among all experts in the study. Therefore, we accepted the computed values of *precision* and *recall* as reliable measures of the structurization pipeline's performance.

High values of *precision* mean that the majority of RTs returned by the structurization process accurately reflect informational points of the original pathology report. High values of *recall* indicate that the majority of the informational points in the report have been carried out to the corresponding KGs in form of RTs. These results demonstrate the pipeline effectively extracts and structurizes complex diagnostic information in free-text pathology reports.

Emerging properties, limitations, and future work

The transformation of diagnostic information from free-text pathology reports into KGs allows linkage of multiple *informational points*. Because of that, the retention of contextual information occurs naturally as a part of the *reification* procedure ("*statement about statement*"). Moreover, comprehensive ontological modeling of the DEs allows for complex and inexact semantical queries. For instance, a query can be constructed to retrieve reports where an IHC study was performed to detect the presence of B-cells. Since, multiple IHC antibodies can be used for this purpose, and their functions are encoded in the DPO, there is no need to run multiple queries for each antibody separately. The system is "smart enough" to recognize the semantics of the query.

As it was discussed in the previous sections, the performance of our structurization pipeline was considered to be sufficient to tackle the task of extracting complex DEs, their relations and structurization of free-text pathology reports for downstream data-mining applications. However, some applications might require higher *recall* values. For instance, in some studies related to Quality Assurance and Quality Control, it could be critical to be able extract all DEs. Missing one or more DE that represents an important diagnostic clue can lead to inconclusive or erroneous results. Since this property depends on the *recall* values of the underlying information extraction technology, we are planning to explore other openIE and non-openIE methods of information extraction.

We have to note here that from a purely technical perspective our structurization pipeline can be viewed as an *ad-hoc* solution, since an extensive vocabulary of terms (DPV) needs to be developed for each pathology practice. However, as we emphasized in the introduction, we believe that this is the only way we can handle variability of expressing complex DEs in narrative text by different diagnostic professionals. Hence, our hypothesis was that only using an extensive DPV, tuned to a specific pathology practice, and AI-based frameworks, enabling computers to act intelligently as humans, we can develop a method to reliably extract complex entities in free-text pathology reports.

Furthermore, AI-based representation of reports enables description logic inference, which can help identify and study implicit relationships among various diagnostic factors. This is the primary goal of our future work.

Conclusion

In this article, we have introduced a novel informatics pipeline to transform free-text diagnostic reports into a structured format. Our work extends openIE techniques with AI-based semantic modeling to extract complex DEs and relationships among them. Evaluation studies have demonstrated that the structurization pipeline possess important properties such as accuracy, *minimality*, and accurate knowledge representation. Therefore, we conclude that our pipeline can be used in various downstream data mining applications in diagnostic medicine such as quality assurance, patient cohort identification, and cancer surveillance.

Acknowledgments

The authors thank the Department of Pathology and Anatomical Sciences, University of Missouri School of Medicine for providing data and resources for this study. We also thank the University of Missouri Institute for Data Science and Informatics for providing support and resources for this study.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Sun R, Medeiros LJ, Young KH. Diagnostic and predictive biomarkers for lymphoma diagnosis and treatment in the era of precision medicine. Mod Pathol 2016;29:1118-42.
2. Higgins RA, Blankenship JE, Kinney MC. Application of immunohistochemistry in the diagnosis of non-Hodgkin and Hodgkin lymphoma. Arch Pathol Lab Med 2008;132:441-61.
3. O'Malley DP, Fedoriw Y, Weiss LM. Distinguishing classical Hodgkin lymphoma, gray zone lymphoma, and large B-cell lymphoma: A proposed scoring system. Appl Immunohistochem Mol Morphol 2016;24:535-40.
4. Murari M, Pandey R. A synoptic reporting system for bone marrow aspiration and core biopsy specimens. Arch Pathol Lab Med 2006;130:1825-9.
5. Camicia R, Winkler HC, Hassa PO. Novel drug targets for personalized precision medicine in relapsed/refractory diffuse large B-cell lymphoma: A comprehensive review. Mol Cancer 2015;14:207.
6. 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;22:1009-19.
7. Martinez D, Li Y. Information extraction from pathology reports in a hospital setting. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York, USA; 2011. p. 1877-82.
8. Bast H, Haussmann E. More informative open information extraction via simple inference. In: de Rijke M. et al. (eds) Advances in Information Retrieval. ECIR 2014. Lecture Notes in Computer Science, Springer, Cham, 2014;8416;585-90.
9. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. N Engl J Med 2018;379:1452-62.
10. Campbell WS, Karlsson D, Vreeman DJ, Lazenby AJ, Talmon GA, Campbell JR. A computable pathology report for precision medicine: Extending an observables ontology unifying SNOMED CT and LOINC. J Am Med Inform Assoc 2018;25:259-66.
11. Foran DJ, Chen W, Chu H, Sadimin E, Loh D, Riedlinger G, *et al*. Roadmap to a comprehensive clinical data warehouse for precision medicine applications in oncology. Cancer Inform 2017;16:1-10.
12. Sarmiento RF, Dernoncourt F. Improving Patient Cohort Identification Using Natural Language Processing. In: Secondary Analysis of Electronic Health Records. Springer, Champ; book chapter; 2016. p. 405-17.
13. Vydiswaran VG, Strayhorn A, Zhao X, Robinson P, Agarwal M, Bagazinski E, *et al*. Hybrid bag of approaches to characterize selection criteria for cohort identification. J Am Med Inform Assoc 2019;26:1172-80.
14. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, *et al*. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA: The Journal of the American Medical Association, 318(22), 2199-210.
15. Shin D, Arthur G, Popescu M, Korkin D, Shyu CR. Uncovering influence links in molecular knowledge networks to streamline personalized medicine. J Biomed Inform 2014;52:394-405.
16. Shin D, Kovalenko M, Ersoy I, Li Y, Doll D, Shyu CR, *et al*. PathEdEx – Uncovering high-explanatory visual diagnostics heuristics using digital pathology and multiscale gaze data. J Pathol Inform 2017;8:29.
17. Al-Taie Z, Thanintorn N, Ersoy I, Kholod O, Taylor K, Hammer R, *et al*. REDESIGN: RDF-based differential signaling framework for precision medicine analytics. AMIA Jt Summits Transl Sci Proc 2018;2017:35-44.
18. He B, Tang J, Ding Y, Wang H, Sun Y, Shin JH, *et al*. Mining relational paths in integrated biomedical data. PLoS One 2011;6:e27506.
19. Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, *et al*. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. J Biomed Inform 2009;42:937-49.
20. Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK, *et al*. The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 2012;3:23.
21. Gao S, Young MT, Qiu JX, Yoon HJ, Christian JB, Fearn PA, *et al*. Hierarchical attention networks for information extraction from cancer pathology reports. J Am Med Inform Assoc 2018;25:321-30.
22. Xie F, Lee J, Munoz-Plaza CE, Hahn EE, Chen W. Application of text information extraction system for real-time cancer case identification in an integrated health care organization. J Pathol Inform 2017;8:48.
23. Yala A, Barzilay R, Salama L, Griffin M, Sollender G, Bardia A, *et al*. Using machine learning to parse breast pathology reports. Breast Cancer Res Treat 2017;161:203-11.
24. Qiu JX, Yoon HJ, Fearn PA, Tourassi GD. Deep learning for automated extraction of primary sites from cancer pathology reports. IEEE J Biomed Health Inform 2018;22:244-51.
25. Napolitano G, Fox C, Middleton R, Connolly D. Pattern-based information extraction from pathology reports for cancer registration. Cancer Causes Control 2010;21:1887-94.
26. Friedman C. Towards a comprehensive medical language processing system: methods and issues. Proc AMIA Annu Fall Symp 1997;595-9.
27. Lee J, Yoon W, Kim S, Kim D, Kim S, So C H, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics; 2019. p. 1-7.
28. Savova GK, Tseytlin E, Finan S, Castine M, Miller T, Medvedeva O, *et al*. DeepPhe: A natural language processing system for extracting cancer phenotypes from clinical records. Cancer Res 2017;77:e115-8.
29. Banko M, Cafarella MJ, Soderland S, Broadhead M, Etzioni O. Open information extraction from the web. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence. San Francisco, CA, USA; 2007. p. 2670-6.
30. Niklaus C, Cetto M, Freitas A, Handschuh S. A survey on open information extraction. In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA; 2018. p. 3866-78.
31. Wu F, Weld DS. Open information extraction using Wikipedia. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA; 2010. p. 118-27.
32. Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA; 2011. p. 1535-45.
33. Mausam M. Open Information Extraction Systems and Downstream Applications. IJCAI'16: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence; AAAI Press; 2016; p. 4074-7; ISBN: 978-1-57735-770-4.
34. Akbik A, Löser A, Krake N. N-ary facts in open information extraction. In: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction (AKBC-WEKEX). Montrèal, Canada; 2012. p. 52-6.
35. Mesquita F, Schmidek J, Barbosa D. Effectiveness and efficiency of open relation extraction. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA; 2013. p. 447-57.
36. Del Corro L, Gemulla R, ClausI E. Clause-based open information extraction. In: Proceedings of the 22Nd International Conference on World Wide Web. New York, NY, USA; 2013. p. 355-66.
37. Angeli G, Premkumar MJ, Manning CD. Leveraging Linguistic Structure for Open Domain Information Extraction. ACL; 2015. p. 344-54.
38. Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA; 2005. p. 363-70.
39. Carvalho JP, Curto S. Fuzzy Preprocessing of Medical Text Annotations of Intensive Care Units Patients | Request PDF. ResearchGate; 2014. Available from: https://www.researchgate.net/publication/286669362_

Fuzzy_preprocessing_of_medical_text_annotations_of_intensive_care_units_patients. [Last accessed on 2019 Feb 21].

40. Gupta A, Banerjee I, Rubin DL. Automatic information extraction from unstructured mammography reports using distributed semantics. J Biomed Inform 2018;78:78-86.
41. Neustein A, Sagar Imambi S, Rodrigues M, Teixeira A, Ferreira L. 1 Application of text mining to biomedical knowledge extraction : Analyzing clinical narratives and medical literature; 2014.
42. The Stanford Natural Language Processing Group, the Stanford Natural Language Processing Group; 2015. Available from: https://nlp.stanford.edu/software/openie.html. [Last Accessed on: 2017 Jun 12].
43. Piskorski J, Yangarber R. Information extraction: Past, present and future. In: Poibeau T, Saggion H, Piskorski J, Yangarber R, editors. Multi-Source, Multilingual Information Extraction and Summarization. Berlin Heidelberg: Springer; 2013. p. 23-49.
44. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: A review of recent research. Geissbuhler A, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2008; Methods Inf Med 2008; 2008; 47 Suppl 1: 128-44 p. 128-144.
45. Zhou L, Plasek JM, Mahoney LM, Karipineni N, Chang F, Yan X, *et al*. Using medical text extraction, reasoning and mapping system (MTERMS) to process medication information in outpatient clinical notes. AMIA Annu Symp Proc 2011;2011:1639-48.
46. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: An introduction. J Am Med Inform Assoc 2011;18:544-51.
47. Hayes P, Carroll J, Welty C, Uschold M, Vatant B, Manola F, *et al*. Defining N-Ary Relations on the Semantic Web. Defining N-ary Relations on the Semantic Web W3C Working Group Note; 12, April 2006. Available from: https://www.w3.org/TR/swbp-naryRelations/. [Last accessed on 2017 Jun 12].
48. Xavier CC, Lima DV, Souza M. Open information extraction based on lexical semantics. J Braz Comput Soc 2015;21:4.
49. Bellogín A, Castells P, Cantador I. Statistical biases in information retrieval metrics for recommender systems. Inf Retr J 2017;20:606-34.
50. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval by Christopher D. Manning. Cambridge Core; 2008. Available from: http/core/books/introduction-to-information-retrieval/669D108D20F556C5C30957D63B5AB65C. [Last accessed on 2019 Jan 10].
51. Roldán É. Dissipation and Kullback-Leibler Divergence. In: Irreversibility and Dissipation in Microscopic Systems; book chapter: 2; Springer Theses; Springer International Publishing Switzerland; 2014; p. 37-59; DOI 10.1007/978-3-319-07079-7.
52. Bigi B. Using Kullback-Leibler distance for text categorization. In: Proceedings of the 25th European Conference on IR Research. Berlin, Heidelberg; 2003. p. 30-19.
53. Wang K, Phillips CA, Saxton AM, Langston MA. EntropyExplorer: An R package for computing and comparing differential Shannon entropy, differential coefficient of variation and differential expression. BMC Res Notes 2015;8:832.
54. Massey FJ Jr. The Kolmogorov-Smirnov Test for Goodness of Fit, Journal of the American Statistical Association, 1951; 46:253; 68-78; DOI: 10.1080/01621459.1951.10500769.
55. Chiodini P, Facchinetti S. Exact critical values of kolmogorov-smirnov test for discrete random variables. Stat Appl 2011;9:63-77.
56. Higgins JJ. Introduction to Modern Nonparametric Statistics. 1st ed. Pacific Grove, CA: Duxbury Press, 2003.
57. Pilehvar MT, Jurgens D, Navigli R. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Long Papers. Vol. 1. Sofia, Bulgaria; 2013. p. 1341-51.
58. Pilehvar MT, Navigli R. An Open-source Framework for Multi-level Semantic Similarity Measurement. Proceedings of NAACL-HLT; 2015.
59. Haveliwala TH. Topic-sensitive pagerank. In: Proceedings of the 11th International Conference on World Wide Web. New York, NY, USA; 2002. p. 517-26.
60. Fellbaum, C. WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics; Second Edition; Oxford: Elsevier; 2005; 665-670; eBook ISBN: 9780080547848.
61. Vickery BC. Techniques of Information Retrieval. London: Butterworths; 1970.
62. Likert R. A technique for the measurement of attitudes. Psychol 1932;22:55.
63. Ludbrook J. Analysis of 2 × 2 tables of frequencies: Matching test to experimental design. Int J Epidemiol 2008;37:1430-5.
64. Warner P. Testing association with fisher's exact test. J Fam Plann Reprod Health Care 2013;39:281-4.
65. Kim HY. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. Restor Dent Endod 2017;42:152-5.
66. Freeman JV, Campbell MJ. The analysis of categorical data: Fisher's exact test. Scope; 2007; 16; p. 11-12; https://www.sheffield.ac.uk/polopoly_fs/1.43998!/file/tutorial-9-fishers.pdf [Last accessed on 2019 Apr 10].
67. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15:155-63.
68. Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. Stat Med 1994;13:2465-76.
69. Tinsley HE, Weiss DJ. Interrater reliability and agreement of subjective judgement. Psychol 1975;22:358-76.
70. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30-46.
71. Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. Tutor Quant Methods Psychol 2012;8:23-34.
72. Cicchett DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6:284-90.
73. McHugh ML. Interrater reliability: The kappa statistic. Biochem Med (Zagreb) 2012;22:276-82.
74. Gamer M, Lemon J, Singh IF. Irr: Various Coefficients of Interrater Reliability and Agreement; 2012.