Modeling of Contextual Information in Knowledge Graphs of Diagnostic Reports

Pericles Giannaris[1], Cynthia Tang[2], Olha Kholod[1], Steve Hanson[2], Chi-Ren Shyu[1], Richard Hammer[2], Dong Xu[3], and Dmitriy Shin[2,1]

[1] *MU Informatics Institute,* [2]*Department of Pathology and Anatomical Sciences,*
[3]*Department of Electrical Engineering and Computer Science, University of Missouri,*
1 Hospital Dr. M251 Pathology, Med Sci Bldg, Columbia, MO, 65203, USA
**Email: shindm@health.missouri.edu*

Abstract. The output of the majority of NLP based informatics pipelines for structurization of free-text lack an ability to recover and convey implicit information, found in diagnostic reports. Such information is readily perceived and taken into account by a human reader as a contextual component. Here, we have developed a method to model contextual information in order to recover implicit relationships among structurized diagnostic entities. Our method enables structurization of contextual information into a cohesive and holistic representation of free-text diagnostic reports, which we call Knowledge Graphs. An expert assessment confirmed the capability of the method to correctly convey contextual information. The precision of matching of the semantical content of the free-text with the corresponding knowledge graphs was 0.92 and the recall was 0.84. The Fisher's exact test had odds ratio 19.7 and p-value of 2.2e-16. The intra-correlation coefficient (ICC) statistic that reflects the level of correlation and magnitude of agreement between domain experts was 0.818 (p-value of 0.99). These results indicate high level of agreement among all experts in the study.

Keywords: Contextual Modeling, Knowledge Graphs, Diagnostic Reports.

1 Introduction

Free-text sections of diagnostic reports contain descriptions of molecular data, microscopic findings from biopsy specimens, interpretations of laboratory values, clues for the identification of diseases, and data on disease surveillance.

In order to computationally analyze diagnostic reports, we need to convert free text to a structured format. In this regard, natural language processing information extraction techniques (NLP-IE) have been widely used to automatically extract knowledge from free text via a structured relational triple format [1]. Relational triples are logical structures in the form of subject-predicate-object statements.

Several NLP-IE applications successfully convert free-text to relational triples. Mausam et al. developed *OLLIE*, which extracted relational triples mediated by verbs, nouns and adjectives [2]; Akbik and Loser developed *KRAKEN*, which extracted *n*-ary relations from sentences based on heuristic rules [3], [4]; Bast and Haussmann developed *CSD-IE*, which extracted relational triples that span over several clauses by decomposing a sentence to sub-sequences that *"semantically belong together"* [5], [6]; and Angeli et al. developed *Stanford OpenIE,* which generated relational triples at multiple levels of granularity by learning a classifier to split a sentence in shorter clauses [7].

Structured representation of text facilitates the use of computational models. Then, computers can be used to mine for implicit relations between the data, to discover patterns in the data, and to enable *"semantic understanding and prompt retrieval"* of specific information from documents [8].

Note here that the majority of these current NLP-IE applications extract explicit relations between entities that belong to the same sentence or clause [1], [4], [5]. Consequently, the information based on implicit relationships across a document is discounted. Consider, for example, the relational triples in Figure 1 that are extracted using a NLP-IE application from the following excerpt from a diagnostic pathology report:

Histologic sections show partial effacement of the lymph node with areas of nodular architecture [. . .]. Scattered Reed- Sternberg cells are present [. . .].

Subject	Predicate	Object
histologic sections	show	effacement of lymph node
histologic sections	show	effacement
histologic sections	show partial effacement	area of nodular architecture
scattered Reed-Sternberg cells	are	present

Figure 1Relational triples generated from sentences in a diagnostic report using NLP-IE applications

According to an expert pathologist, the focus of the diagnostic report is a *"lymph node"*. In this case, *Reed-Sternberg* cells should be considered in the context of that *"lymph node"*. However, since this fact was not expressed explicitly in the text, structurization algorithms would not convey it in their output.

As illustrated from the previous examples, context is essential component of NLP-IE. In biomedicine, the term "context" describes entities related to a biomedical problem. Although the context of a whole text document is essential for the extraction of implicit information, most NLP-IE applications instead focus on the context of a sentence or a clause. For example, Mausam et al. use *"attribution and clausal modifiers"* to extend a relational triple to a quadruple. The extra field provides contextual information [5] [2], this process is also known as *reification*. Bast and Haussmann based on constituent parsing, which splits a sentence to parts that *"semantically belong together, [to form] so-called 'contexts'"*. Here, each *"context"* is a fact that depends on surrounding *"contexts"* [5]. Similarly, Angeli et al. use *natural logic annotations* to split a sentence to shorter clauses thus, enabling the *"system to have a greater awareness of the context of each extraction"* by generating multiple instances of the same relation [7]. For this study, we use the *Stanford OpenIE* application by Angeli et al. to extract information within the context of the following excerpt:

Histologic sections show partial effacement of the lymph node with areas of nodular architecture. The nodular areas are composed of a mixture of monomorphic small cells [. . .]. Immunostains with the appropriate controls are performed on block 1A [. . .]. Reed-Sternberg cells mark with weak nuclear positivity for PAX5 [. . .]. Reed-Sternberg cells are negative for CD20.

Subject	Predicate	Object
histologic sections	show	partial effacement of lymph node with areas of nodular architecture
nodular areas	are composed of	mixture of monomorphic small cells
immunostains	are performed on	block-1A
Reed-Sternberg cells	mark with	weak nuclear positivity for PAX5
Reed-Sternberg cells	are negative for	CD20

Figure 2 Example of information extracted in the context of a diagnostic report using openIE applications.

The relational triples in Figure 2 demonstrate that the current NLP-IE system has *"awareness of the context of each extraction"* [7]. This means that it generates triples only from the input sentence. However, implicit relationships between *CD20*, *block 1A*, and *Reed-Sternberg cells* within the context of the *lymph node* are not captured by the NLP-IE system. For example, it is critical for a pathologist to know that *CD20* is negative in *block 1A* for the *lymph node.*

In order to structurize implicit information, we need to model the context of a diagnostic report in relational triple resource description framework *RDF*-like format, which is a building block of a knowledge base (KB). Triples that share *subject* or *object* induce a graph that we link using the *n*-ary relation schema according to the semantic web [9]. We define these graphs as knowledge graphs (KG).

The following section discusses in detail our methodological approach.

2 Methods

The informatics pipeline for modeling of contextual information is implemented as two independent processes (See Figure 3). The following sections describe these processes in detail.

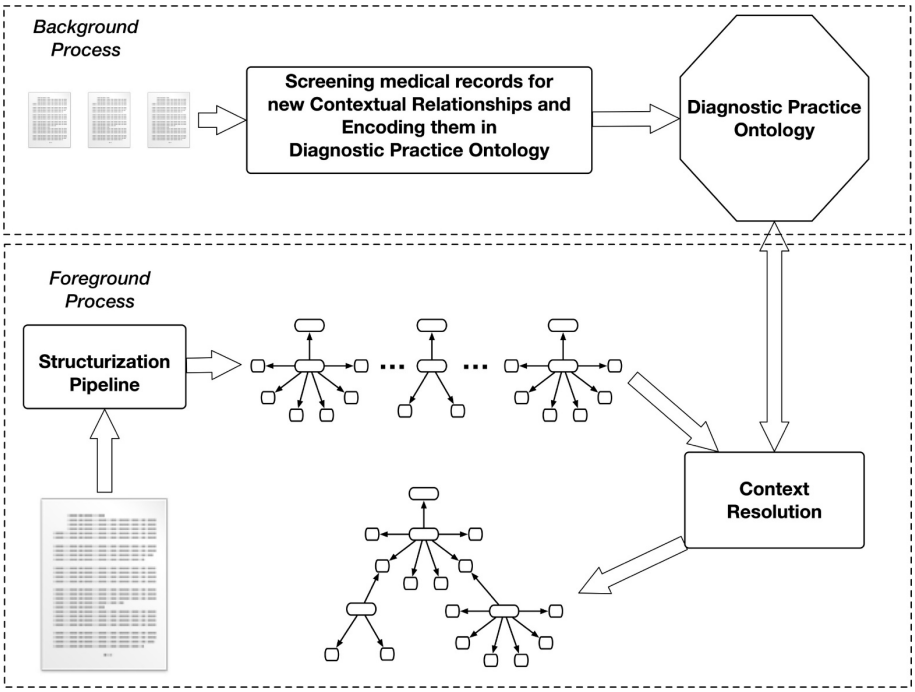

Figure 3. Contextual modeling framework is implemented as two independent processes.

2.1 Contextual Encoding

To encode contextual information, we use a Diagnostic Practice Ontology (DPO). DPO consists of concepts and relationships that describe a specific diagnostic setting. For instance, DPO includes concepts related to the diagnostic process such as types of specimens, tissues and cells as well as various diagnostic tests. It also includes a hierarchy of personnel involved in a diagnostic process such as pathologists, residents, and laboratory staff. A structurized version of a diagnostic report consists of instantiations of DPO concepts and their relationships.

In a semi-automatic *Background Process* (top panel in Figure 3), diagnostic reports from a Laboratory Information System (LIS) are screened by a human expert to search for new contextual relationships that are not present in the DPO. For instance, headings of sections of diagnostic reports can represent contexts for concepts described in these sections. Another example is specific locations in tissue specimens that can serve as context for biological entities (e.g. germinal center is context for Reed-Stenberg cells).

To encode these relationships, we introduce a notion of *Contextual Ancestry* (CA). Given a concept *A* from a diagnostic report (e.g. a diagnostic test, a molecular entity, a specific cell), a *Contextual Ancestry* represents diagnostic concepts that can serve as a *context* for the concept *A*. Such *contextual concepts* are arranged in the CA in the order of their appearance in the reports. For instance, CD4 immunohistochemical (IHC) antibody test can have the following CA:

CD4->Block 1A->IHC study->Lymph Node->Surgical Report

Diagnostic concepts Block 1A, IHC study, Lymph Node, Surgical Report can all serve as a context for CD4 IHC test. We have to note here, that *Contextual Ancestries* are specific to a LIS platform used in a pathology practice, in the sense that they reflect the style and order of sections of a diagnostic report generated by that LIS. Contextual Ancestries can form hierarchical structures like trees or even networks. The above example represent a linear path in the CA hierarchy.

The generated CAs are then incorporated in real time into a Diagnostic Practice Ontology (DPO), which models various types of relationships required to structurize diagnostic reports.

2.2 Context Resolution

Structurization of implicit diagnostic information is performed through the *Context Resolution* step of the pipeline in an automatic fashion (bottom panel in Figure 3). First, a diagnostic report is processed by a structurization pipeline to generate relational triples. To do this, we utilize Stanford OpenIE software. The resulting RDF triples are then arranged as *n-ary relation* models. Such *n-ary* models represent structurized version of specific informational points from the diagnostic report. However, while they may successfully convey the intended semantics, in many cases, they lack the contextual component. The lack of contextual information may undermine the usefulness of the *n-ary* models. To demonstrate this, consider the following excerpt from a diagnostic pathology report (Figure 4):

Surgical Report: S17-0987
Patient: Joe Doe
Accession Date: 04/23/2017
MRN: 0001

Histologic sections show partial effacement of the lymph node with areas of nodular architecture. The nodular areas are composed of a mixture of monomorphic small cells and fewer intermediate large enlarged cells. Scattered Reed-Sternberg cells are present. There is vascular proliferation and prominent deposition of pink dysproteinemic material in the center of nodules with accompanying fibrosis. The areas of residual follicles have prominent surrounding plasmacytosis. Throughout the sections there is perivascular fibrosis with "onion skinning" around the vessels. There is also capsular fibrosis noted. Increased numbers of eosinophils are not seen. Immunostains with the appropriate controls are performed on block 1A and show Reed-Sternberg cells are positive for CD30 and have membrane and Golgi positivity for CD15. Reed-Sternberg cells mark with weak nuclear positivity for PAX5. Reed-Sternberg cells are negative for CD20, CD3, CD43, BCL6, CD79a, EMA, Alk-1, and LCA (CD45).

Figure 4. Excerpt form a diagnostic pathology report.

The report includes a statement about the presence of Reed-Stenberg (RS) cells. The corresponding structurized *n-ary* model (*Microscopic_Description_002*) that conveys

this information is shown at the right bottom of the upper panel in Figure 5. It can be clearly seen that the structurized version of the report does not include information in which type of tissue RS cells were spotted. This type of critical information is implicitly conveyed by the first sentence in the excerpt (underlined in Figure 4) and noted by a human reader. Other examples include implicit contextual relationships between *IHC_Study_001* and *Surgical_Report_001*, *IHC_Study_002* and *Surgical_Report_001*, *Microscopic_Description_001* and *Surgical_Report_001,* and *Microscopic_Description_002* and *Surgical_Report_001*, which connect specific microscopic findings and test results to patient information. The implicitly reported location of RS cells after context resolution step is properly established as being in lymphoid tissue. The generated contextual relationships are represented by a RDF predicate *in_context_of* and marked with purple in Figure 5.

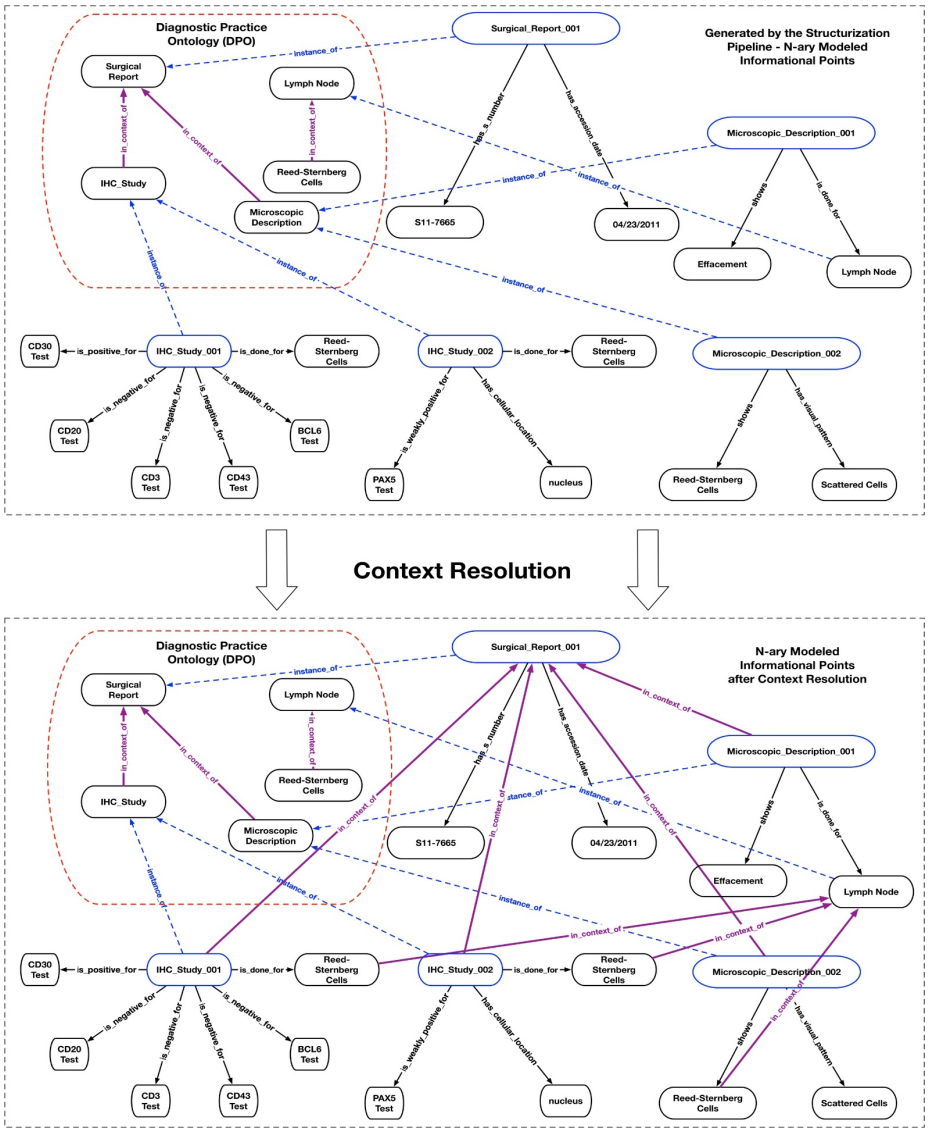

Figure 5. An example of Context Resolution to model implicit information in a diagnostic pathology report.

The process of generation of contextual links is depicted by a pseudo-code in Figure 6, which represents an (unoptimized) iterative process of finding context entities for each node in the structurized version of a diagnostic report. We call the overall output of the developed pipeline a *Knowledge Graph*, where all individual structurized informational points are connected in a cohesive manner to holistically represent a diagnostic pathology report.

Procedure **ContextResolution()**

Input:

a. *Diagnostic Domain Ontology*, O = (C, R), where C = {c_1, c_2, … , c_p} is set of *p* diagnostic concepts c_k, and R = {r_1, r_2, … , r_s} is a set of *s* relations r_t between concepts c_k, such that ∃ r_c ∈ R, a contextual relation between concepts c_k

b. A set NM = (nm) of n-ary models nm representing a structurized version of a diagnostic report DR, with each n-ary model nm = {n-ary_anchor, n_1, n_2… n_m}

Output:

A *Knowledge Graph* K_{DR} = (O, R, NM)

1. For each n-ary model nm_x ∈ NM do:
2. For each node n_x ∈ nm_x do:
3. ca_x = getNearestContextualAncestor (n_x, O)
4. If ca_x = null, then Continue Loop (2)
5. Else, do:
6. For each nm_y ∈ NM\nm_x do:
7. For each node n_y ∈ nm_y do:
8. If n_y = ca_x, then do:
9. Connect n_x and n_y using r_c
10. Else, Continue Loop (7)
11. End of Loop (7)
12. End of Loop (6)
13. End of Loop (2)
14. End of Loop (1)

15. Return K_{DR}

Figure 6. A pseudocode of Context Resolution procedure. Loop (2) searches for a contextual ancestor for each node of all n-ary models of the structurized version of a diagnostic report. For that it uses the function getNearestContextualAncerstor(), which traverses Contextual Ancestry tree upwards. If such contextual ancestor is found, Loop (7) checks whether an instantiation of this ancestor is present in any other n-ary model. If that is the case, the node is connected to the instantiation of the contextual ancestor by a contextual relation.

3 Results and Discussion

We analyzed 34 pathology reports that yielded over 3,500 *RDF-like* relational triples that we represented as KGs. We have performed an expert assessment of the effectiveness of conveying implicit contextual information into the generated knowledge graphs of diagnostic reports. For that, we recruited 6 domain experts from the University of Missouri Department of Pathology to evaluate the output of our method. The evaluation was based on three levels of a Likert-like scale: *"in context"*, *"not in context"*, *"not clear"*. We measured the effectiveness of our model with performance statistics for information retrieval systems. The precision metrics of matching the semantical content of the free-text with the corresponding knowledge graphs was 0.92 and the recall was

0.84. A Fisher's exact test was used to assess statistical significance. The Fisher's exact test has odds ratio 19.7 and p-value of 2.2e-16. Inter-Raters' Reliability (IRR) score according to a *two-way random effects model based on a fully crossed design* as described in [10]. The intra-correlation coefficient (ICC) statistic that reflects the level of correlation and magnitude of agreement between domain experts [11] was 0.818 (p-value of 0.99). We, therefore, concluded that the differences in the assessment were *statistically insignificant*. These results indicate high level of agreement among all experts in the study. Therefore, we accepted the computed values of precision and recall as measures of the structurization pipeline's performance. Usage of KGs have several advantages: they represent domain knowledge and facts, they are human and machine readable, and they enable graph mining to discover non-trivial patterns in the data.

4 Conclusion

We have developed a method to model contextual information in order to recover implicit relationships among structurized diagnostic entities. The method enables structurization pipelines to convey contextual information and connect structurized informational point into a cohesive and holistic representation of free-text diagnostic reports, which we call *Knowledge Graphs*. A limitation of our study is the sample size. Our future efforts will concentrate on applying rules to our ontology. Preliminarily, KGs are important in healthcare for data mining aspects and knowledge acquisition.

References

1. J. Piskorski and R. Yangarber, "Information Extraction: Past, Present and Future," in Multi-source, Multilingual Information Extraction and Summarization, T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, Eds. Springer Berlin Heidelberg, 2013, pp. 23–49.
2. Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni, "Open language learning for information extraction," pp. 523–534, Jul. 2012.
3. A. Akbik and A. Löser, "KrakeN: N-ary facts in open information extraction," pp. 52–56, Jun. 2012.
4. C. C. Xavier, V. L. S. de Lima, and M. Souza, "Open information extraction based on lexical semantics," Journal of the Brazilian Computer Society 2015 21:1, vol. 21, no. 1, p. 4, Dec. 2015.
5. C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, "A Survey on Open Information Extraction," presented at the International Conference on Computational Linguistic, Santa Fe, USA, 2018, pp. 3866–3878.
6. H. Bast and E. Haussmann, "Open Information Extraction via Contextual Sentence Decomposition," presented at the 2013 IEEE Seventh International Conference on Semantic Computing (ICSC), pp. 154–159.
7. G. Angeli, M. Premkumar, C. M. O. T. 5. A. M. of, 2015, "Leveraging linguistic structure for open domain information extraction," presented at the rd Annual Meeting of the Association for Computational Linguistics and the th International Joint Conference on Natural Language Processing, Beijing, China, 2015.

8. S. Zheng, J. J. Lu, C. Appin, D. Brat, and F. Wang, “Support patient search on pathology reports with interactive online learning based data extraction,” J. Pathol. Inform., vol. 6, p. 51, 2015.
9. P. Hayes 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, 12-Apr-2006. [Online]. Available: https://www.w3.org/TR/swbp-n-aryRelations/. [Accessed: 12-Jun-2017].
10. K. A. Hallgren, “Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial,” Tutor Quant Methods Psychol, vol. 8, no. 1, pp. 23–34, 2012.
11. T. K. Koo and M. Y. Li, “A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research,” Journal of Chiropractic Medicine, vol. 15, no. 2, pp. 155–163, Jun. 2016.