Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2016 February ; 2016: 449–452. doi:10.1109/BHI.2016.7455931.

Toward patient-tailored summarization of lung cancer literature*

Jean I. Garcia-Gathright[1], **Nicholas J. Matiasz**[1], **Edward B. Garon**[2], **Denise R. Aberle**[1,3], **Ricky K. Taira**[1,3], and **Alex A. T. Bui**[1,3]

[1]University of California Los Angeles, Department of Bioengineering

[2]University of California Los Angeles, Department of Medicine

[3]University of California Los Angeles, Department of Radiological Sciences

Abstract

As the volume of biomedical literature increases, it can be challenging for clinicians to stay up-to-date. Graphical summarization systems help by condensing knowledge into networks of entities and relations. However, existing systems present relations out of context, ignoring key details such as study population. To better support precision medicine, summarization systems should include such information to contextualize and tailor results to individual patients.

This paper introduces “contextualized semantic maps” for patient-tailored graphical summarization of published literature. These efforts are demonstrated in the domain of driver mutations in non-small cell lung cancer (NSCLC). A representation for relations and study population context in NSCLC was developed. An annotated gold standard for this representation was created from a set of 135 abstracts; F1-score annotator agreement was 0.78 for context and 0.68 for relations. Visualizing the contextualized relations demonstrated that context facilitates the discovery of key findings that are relevant to patient-oriented queries.

I. INTRODUCTION

The domain of non-small cell lung cancer (NSCLC) is fertile ground for new developments in precision oncology. The Lung Cancer Mutation Consortium recently characterized ten driver mutations in lung cancer, and the Federal Drug Administration has approved targeted therapies for patients with epidermal growth factor receptor (EGFR) mutations and anaplastic lymphoma kinase (ALK) rearrangements [1]. Improvement of clinical outcomes with targeted therapies has been demonstrated in clinical trials, and new advances continue to be made [2].

Biomedical literature is an abundant source of knowledge, but clinicians have a limited amount of time to review the literature. Summarization systems help by presenting knowledge in a condensed form. Furthermore, in accordance with the efforts of precision medicine, a summarization system should be capable of providing information that is

*This work was supported by NLM T15-LM007356, NIH/NLM R01-LM009961, NIH T32-EB016640-02, and the UCLA Department of Radiological Sciences.

Corresponding author: jigarcia@ucla.edu.

tailored to a specific patient's characteristics, such as mutation status, treatment history, and mechanism of resistance.

One approach to summarization involves representing scientific claims as propositions, such as "chemotherapy improves survival." These propositions can then be represented graphically, as nodes ("chemotherapy" and "survival") and edges signifying a relation between nodes ("improves"). However, existing graphical summarization systems depict these relations out of context, ignoring key information such as study population characteristics. To judge a relation's applicability to a specific patient and capture its full meaning, these contexts are crucial.

This paper describes the "contextualized semantic map," a representation that ties relations to their associated contexts. This approach to graphical summarization is demonstrated in the domain of driver mutations in NSCLC.

II. BACKGROUND

The idea of representing propositional knowledge as nodes and edges in a graph was first introduced by Novak in 1977 [3]. Since then, many in the biomedical domain have used relations to summarize knowledge about protein-protein interactions [4], [5], gene-protein interactions [6], [7], and relationships between treatments and diseases [8], [9]. Notably, SemRep extracts concepts and relations from the Unified Medical Language System [10] to summarize biomedical journal articles [11].

While the artificial intelligence community has long recognized the importance of context in knowledge-based systems [12], few biomedically-oriented relation extraction systems fully leverage context in their information retrieval and summarizations. BioContext is a text mining system that contextualizes biomolecular events in terms of species, anatomical location, speculation, and negation [13]. BIO-SMILE augments relations with the surrounding words signifying the location, manner, and timing of an event [14]. Semantic MEDLINE, a biomedical graphical summarization system based on SemRep relations, uses statistical features (such as frequency of occurrence in the corpus) and graph-based features (such as adjacency to a node of interest) to focus its graphical summary [15]. The PICO representation leverages study population for information retrieval and question answering, but not for summarization or relation contextualization [16].

This paper describes the development and use of a detailed representation of study population for the purpose of contextualizing relations in the domain of non-small cell lung cancer. The contributions of this work are: 1) a semantic approach to graphical summarization that includes contextual information; 2) an annotated gold standard of relations and study population context in NSCLC; and 3) a first pass at visualization of contextualized relations.

III. METHODS

A. Representation

1) Study population context—We first define the representation for study population context. This representation was based on the National Lung Cancer Audit (LUCADA), an effort in the United Kingdom to create a registry of lung cancer patients and their treatments and outcomes [17]. The LUCADA representation includes information on patient demographics, risk factors, treatment history, and tumor features. The representation was augmented based on expert opinion (EG, DA) to include information on driver mutations, targeted therapy, imaging features, and clinical response. An overview of the representation is given in Figure 1.

2) Targeted concepts and relations—Next, we define the sets of targeted concepts and relations of interest. In our previous work, we identified four common study objectives: mutation characterization, mutation detection, treatment, and prognosis [18]. These study objectives are used to develop a set of targeted relations. Definitions of targeted concepts are described in Table I. When possible, definitions from resources such as the National Cancer Institute Thesaurus [19] or Unified Medical Language System [10] were used. Relations appropriate for each study objective are given in Table II. Any relation may be augmented with study population context, indicating that the relation was found in a population possessing certain clinical features.

B. Data collection and annotation

To validate the coverage and specificity of the representation, an annotated gold standard of relations and concepts was created manually from a set of abstracts on EGFR mutation in lung cancer. These abstracts were obtained from two sources: a snapshot of PubMed and archives from the annual meeting of the American Society of Clinical Oncologists. In a previous study, we searched these two sources for articles published in 2013 containing “EGFR” and “lung” in the title, resulting in 157 abstracts [18]. Studies identified as out of scope (e.g., case reports, pre-clinical studies) were excluded, resulting in a total of 135 abstracts.

Annotation guidelines were created to formalize the annotation of relations and population context. The annotation guidelines provided definitions of each concept and relation, as well as specific directions regarding the scope of the annotations. For example, study population context is limited to eligibility criteria rather than descriptive statistics of the cohort. Relations of interest expressing the findings of the study were limited to the Results and Conclusion sections of the abstracts.

Mentions of concepts and relations within the abstracts were annotated using brat rapid annotation tool[1] by the lead author (JG) and a biomedical informatics graduate student (NJM). The readers annotated the corpus independently; agreement was then calculated, and entities with low agreement were selected for discussion. After discussing several points of

[1] http://brat.nlplab.org

clarification, annotation guidelines were updated. The readers corrected their annotations per the updated guidelines until consensus was achieved.

Annotator agreement was calculated in terms of F1-score, holding JG's annotations as the ground truth for the purposes of evaluation [20]. Agreement was calculated for each relation as well as semantically related groups of relations (e.g., *improves* and *associated with*, *predicts better* and *predicts*, *positive correlation* and *correlation*).

The resulting gold standard is publicly available at http://jigarcia.bol.ucla.edu.

C. Visualization

A contextualized semantic map was produced by loading the manually-annotated relations and contexts into a graph structure using the Python library networkx[2]. Each relation (edge) has a set of attribute-value pairs corresponding to the concepts in the representation and their instances from the annotated document set. The user may filter the network according to population attributes of interest. In this evaluation, we visualize networks produced by two filters: *targeted therapy history*=EGFR-tyrosine kinase inhibitors (TKIs) and *biomarker*=EGFR+. These filters were chosen to simulate queries provided by a clinical expert (EG).

IV. RESULTS

A. Annotator agreement

1) Study population context—Ten concepts in our representation collectively contributed to over 90% of the total annotations. Among these, F1-score agreement ranged from 0.56 for *progression* to 0.92 for *surgery history*. F1-score agreement over all concept types was 0.78.

2) Relations—Thirteen relations collectively contributed to roughly 90% of the total annotations. F1-score agreement for these relations ranged from 0.40 for *associated with* to 0.80 for *does not predict* and *predicts worse*. Overall agreement for relations was 0.68. Combining semantically similar relations proved beneficial to F1-score agreement. Notably, the combined relation *improves or associated with* had an F1-score of 0.79.

Table III presents the agreement values for study population context and relations.

B. Visualization

The resulting contextualized semantic map contains 570 nodes and 591 edges. Filtering the graph by study population context reduces the size of the graph significantly, facilitating discovery of key findings that are relevant to a patient-oriented query. For example, filtering on *targeted therapy history*=EGFR-TKIs produces a subgraph pictured in Figure 2a. In pre-treated populations, TKI resistance is associated with with poor outcomes; however, one study showed that TKIs used with chemotherapy or radiotherapy yields an improvement in overall survival.

[2] http://networkx.github.io

Figure 2b depicts a subgraph depicting a set of *improves* relations. A number of chemotherapies and targeted therapies are identified; however, after applying a filter to identify relations exclusively from studies on EGFR+ cohorts, a smaller number of treatment-oriented relations appears (outlined in bold). Thus, the summary includes information specific to the user's query, including relations only from studies on EGFR+ populations.

V. DISCUSSION

The substantial inter-annotator agreement on the majority of concept and relation types validates the precision of our representation and the suitability of our annotations as a gold standard for development of automatic extraction systems. Furthermore, while the study population representation is specific to lung cancer, many of the features could be applied to cancer in general, opening the possibility of contextualized semantic maps in other cancer domains. Capturing numeric data such as effect sizes and confidence intervals remains an open issue for future work.

We also developed a technique for visualizing contextualized relations, which includes an edge-filtering mechanism that enables the user to view studies relevant to specific patient characteristics. Future improvements to the visualization include vocabulary standardization (i.e., combining synonymous nodes) and semantic clustering (i.e., edges with similar contexts are placed near each other).

Development of a representation for relations and their associated study population context is a first step toward informing clinical decisions through patient-tailored summarization. Ultimately, our goal is to create an end-to-end summarization system, including automatic extraction of relations and context, and evaluation of our representation on a clinical information retrieval task.

References

1. Kris MG, Johnson BE, Berry LD, et al. Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. JAMA. 2014; 311(19):1998–2006. [PubMed: 24846037]
2. Chen Z, Fillmore CM, Hammerman PS, et al. Non-small-cell lung cancers: a heterogeneous set of diseases. Nat Rev Cancer. 2014; 8:535–546.
3. Novak, JD. A theory of education. Ithaca, NY: Cornell University Press; 1977.
4. Ono T, Hishigaki H, Tanigami A, et al. Automated extraction of information on proteinprotein interactions from the biological literature. J Bioinform. 2001; 17(2):155–161.
5. Mooney, RJ., Bunescu, RC. Subsequence kernels for relation extraction. In: Weiss, Y.Schölkopf, B., Platt, J., editors. Advances in Neural Information Processing Systems 18. Cambridge, MA: MIT Press; 1994. p. 171178
6. Fundel K, Küffner R, Zimmer R. RelEx — Relation extraction using dependency parse trees. J Bioinform. 2007; 23(3):365–371.
7. Giuliano, C., Lavelli, A., Romano, L. Exploiting shallow linguistic information for relation extraction from biomedical literature. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06); Trento, Italy. 2006. p. 401408
8. Rosario, B., Hearst, MA. Classifying semantic relations in bioscience texts. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics; Barcelona, Spain. 2004. p. 430

9. Bundschus M, Dejori M, Stetter M, et al. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics. 2008; 9(1):207. [PubMed: 18433469]
10. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004; 32(suppl 1):D267–D270. [PubMed: 14681409]
11. Rindflesch, TC., Fiszman, M., Libbus, B. Semantic interpretation for the biomedical literature. In: Chen, H.Fuller, S.Hersh, WR., Friedman, C., editors. Medical informatics: Advances in knowledge management and data mining in biomedicine. Springer-Verlag; 2005. p. 399422
12. McCarthy, J. Notes on formalizing context. 1993.
13. Gerner M, Sarafraz F, Bergman CM, et al. BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. J Bioinform. 2012; 28(16):2154–2161.
14. Tsai RTH, Chou WC, Su YS, et al. BIOSMILE: a semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. BMC Bioinformatics. 2007; 8(1):325. [PubMed: 17764570]
15. Rindflesch TC, Kilicoglu H, Fiszman M, et al. Semantic MEDLINE: An advanced information management application for biomedicine. Information Services & Use. 2001; 31(1–2):15–21.
16. Demner-Fushman D, Lin J. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. Computational Linguistics. 2007; 33(1):63–103.
17. Rich AL, Tata LJ, Stanley RA, et al. Lung cancer in England: information from the National Lung Cancer Audit (LUCADA). Lung Cancer. 2011; 72(1):16–22. [PubMed: 20688413]
18. Garcia-Gathright JI, Oh A, Abarca PA, et al. Representing and extracting lung cancer study metadata: Study objective and study design. Comput Biol Med. 2015; 58:63–72. [PubMed: 25618216]
19. Sioutos N, de Coronado S, Haber MW, et al. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007; 40(1):30–43. [PubMed: 16697710]
20. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc. 2005; 12(3):296–298. [PubMed: 15684123]

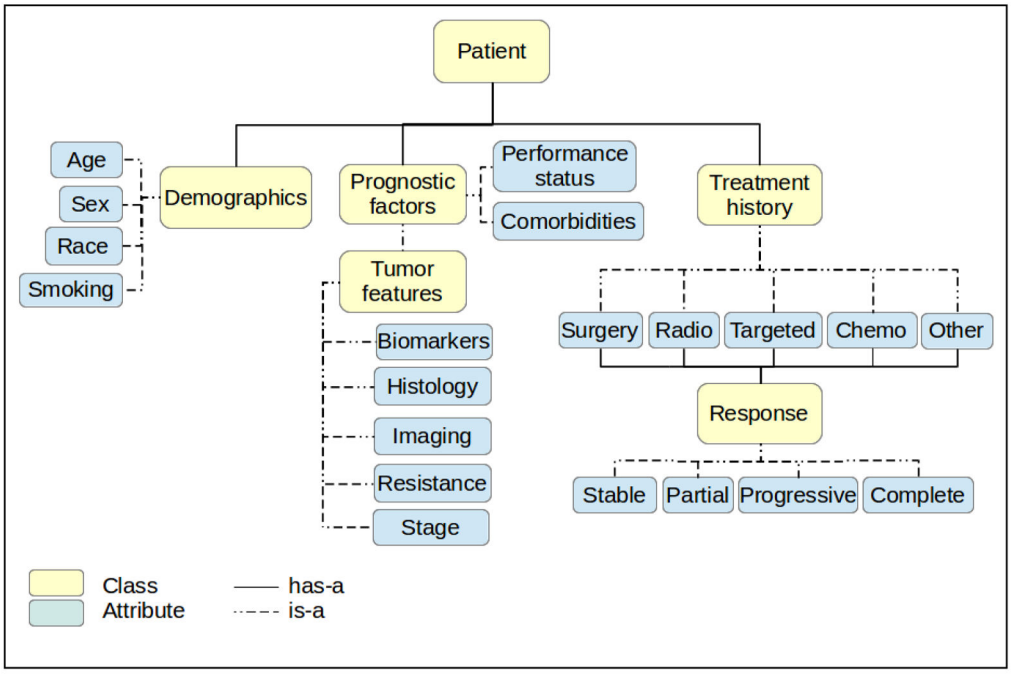

Fig. 1.
A representation for study population context.

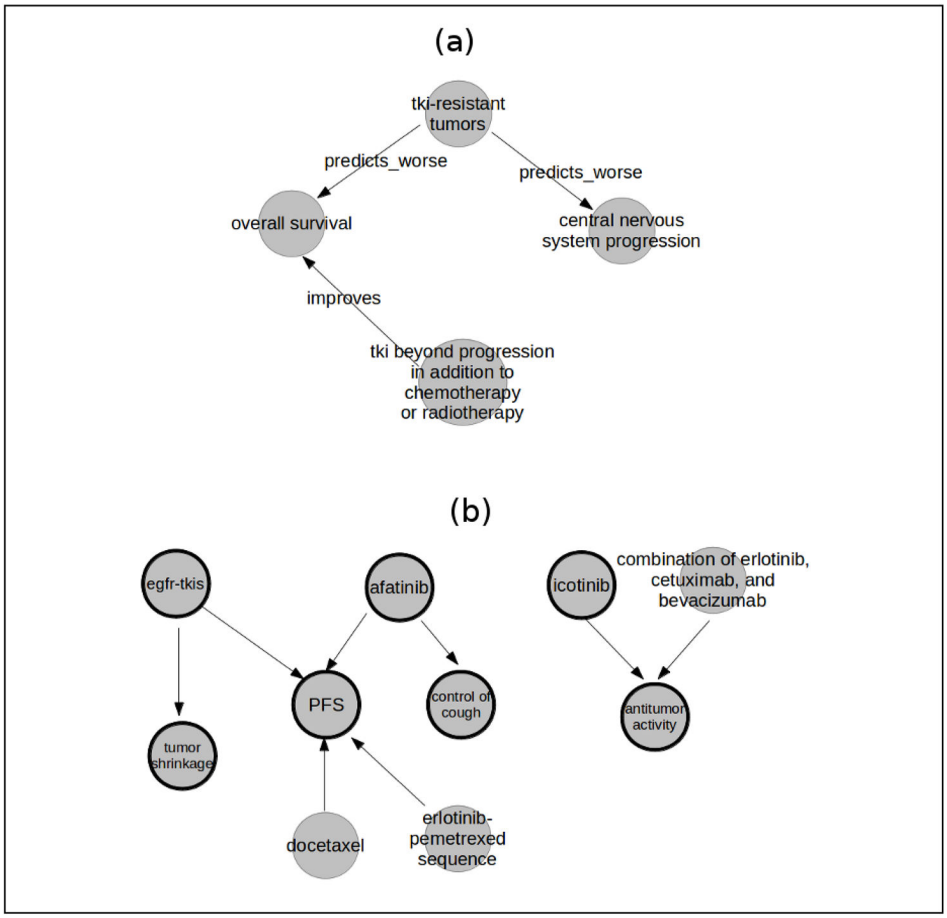

Fig. 2.

Manually diagrammed depictions of contextualized semantic maps.

(a) The contextualized semantic map after filtering for "*targeted therapy history* = EGFR TKIs." The graph has been significantly reduced in size, facilitating knowledge discovery.

(b) A fragment of *improves* relations. Nodes outlined in bold indicate that the associated relations were discovered in a cohort of EGFR+ patients.

TABLE I

Definitions of concepts that participate in relations.

Concept	Definition
Biomarker	A characteristic that serves as an indicator for normal biologic processes, pathogenic processes, state of health or disease, the risk for disease development and/or prognosis, or responsiveness to a therapeutic intervention.[1]
Clinical feature	Clinical-pathologic features of a patient, see Figure 1.[2]
Detection method	A procedure, method, or technique used to determine the nature or identity of a disease or disorder.[3]
Material	An aggregation of similarly specialized cells and the associated intercellular substance.[3]
Outcome	A place of termination or completion, such as a primary or secondary outcome variable used to judge the effectiveness of a treatment.[1]
Treatment	Any type of intervention intended to treat a condition in a patient, including targeted therapy, chemotherapy, radiotherapy, and surgery.[1,3]
Rate	The ratio of the number of occurrences of a disease or event to the number of units at risk in the population.[1]

[1]Source: National Cancer Institute.

[2]Source: National Lung Cancer Audit.

[3]Source: Unified Medical Language System.

TABLE II

List of permitted relations, organized by study objective.

Study objective	Description	Relation
Mutation characterization	Studies that correlate clinical-pathologic features (e.g., age, race, smoking status) with biomarker status, or report the prevalence of mutation within a population or in comparison to another population.	biomarker {**positively, negatively, not**} **correlated with** clinical feature biomarker **has rate** rate biomarker **has** {**higher, lower, similar**} **rate in** clinical feature
Mutation detection	Mutation detection studies demonstrate a method for detecting mutation status, sometimes specifying the type of biological specimen used.	detection method **detects** biomarker biomarker **detected in** material
Treatment	Treatment studies examine the association between treatments and outcomes. Treatments can improve outcomes (e.g., longer survival), worsen outcomes (e.g., side effects), or have no effect on outcomes. Treatments may also be recommended for a specific sub-population.	treatment {**improves, worsens, does not improve, associated with**} outcome treatment **recommended for** clinical feature
Prognosis	Prognosis studies associate clinical-pathologic features and biomarkers with outcomes.	{biomarker, clinical feature} {**predicts, predict better, predicts worse, does not predict**} outcome

TABLE III

Annotator agreement for each concept and relation type.

Concept	Total	Precision	Recall	F-1
biomarker	151	0.68	0.88	0.77
stage	146	0.80	0.85	0.82
histology	93	0.75	0.90	0.82
targeted therapy history	91	0.78	0.80	0.79
ethnicity nationality	67	1.0	0.63	0.78
chemotherapy history	40	0.94	0.77	0.85
progression	32	0.53	0.60	0.56
other treatment history	29	0.56	0.91	0.69
resistance	27	0.71	0.77	0.74
surgery history	26	0.86	1.0	0.92
All concepts	759	0.75	0.81	0.78
Relation				
has rate	207	0.60	0.90	0.72
improves	131	0.88	0.70	0.78
does not predict	113	0.87	0.74	0.80
predicts better	112	0.82	0.66	0.73
predicts worse	112	0.92	0.71	0.80
predicts	107	0.61	0.67	0.64
detects	94	0.66	0.58	0.62
positive correlation	93	0.80	0.63	0.71
detected in	56	0.67	0.77	0.71
correlation	47	0.48	0.78	0.60
associated with	45	0.33	0.50	0.40
has higher rate in	41	0.87	0.50	0.63
recommended for	25	0.75	0.69	0.72
All relations	1320	0.68	0.68	0.68
Combined relations				
predicts better, predicts	215	0.82	0.80	0.81
improves, associated with	173	0.80	0.77	0.79
positive correlation, correlation	140	0.84	0.84	0.84