Computers in Biology and Medicine 92 (2018) 55–63

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Evaluating Casama: Contextualized semantic maps for summarization of lung cancer studies

Jean I. Garcia-Gathright [a,*], Nicholas J. Matiasz [a], Carlos Adame [c], Karthik V. Sarma [a], Lauren Sauer [c], Nova F. Smedley [a], Marshall L. Spiegel [c], Jennifer Strunck [c], Edward B. Garon [c], Ricky K. Taira [a,b], Denise R. Aberle [a,b], Alex A.T. Bui [a,b]

[a] University of California, Los Angeles, Department of Bioengineering, 924 Westwood Boulevard, Suite 420, Los Angeles, CA, 90024, USA
[b] University of California, Los Angeles, Department of Radiological Sciences, 924 Westwood Boulevard, Suite 420, Los Angeles, CA, 90024, USA
[c] University of California, Los Angeles, Department of Medicine - Division of Hematology-Oncology, 924 Westwood Boulevard, Suite 200, Los Angeles, CA, 90024, USA

ARTICLE INFO

Keywords:
Automatic summarization
Knowledge representation
Evaluation

ABSTRACT

Objective: It is crucial for clinicians to stay up to date on current literature in order to apply recent evidence to clinical decision making. Automatic summarization systems can help clinicians quickly view an aggregated summary of literature on a topic. Casama, a representation and summarization system based on "contextualized semantic maps," captures the findings of biomedical studies as well as the contexts associated with patient population and study design. This paper presents a user-oriented evaluation of Casama in comparison to a context-free representation, SemRep.

Materials and methods: The effectiveness of the representation was evaluated by presenting users with manually annotated Casama and SemRep summaries of ten articles on driver mutations in cancer. Automatic annotations were evaluated on a collection of articles on *EGFR* mutation in lung cancer. Seven users completed a questionnaire rating the summarization quality for various topics and applications.

Results: Casama had higher median scores than SemRep for the majority of the topics ($p \leq 0.00032$), all of the applications ($p \leq 0.00089$), and in overall summarization quality ($p \leq 1.5e\text{-}05$). Casama's manual annotations outperformed Casama's automatic annotations ($p = 0.00061$).

Discussion: Casama performed particularly well in the representation of strength of evidence, which was highly rated both quantitatively and qualitatively. Users noted that Casama's less granular, more targeted representation improved usability compared to SemRep.

Conclusion: This evaluation demonstrated the benefits of a contextualized representation for summarizing biomedical literature on cancer. Iteration on specific areas of Casama's representation, further development of its algorithms, and a clinically-oriented evaluation are warranted.

1. Objective

As the volume of published biomedical literature increases at an unprecedented rate, it is challenging for a clinician to stay up to date. Aggregating and summarizing the current state of knowledge in a disease domain can help inform a clinician's thinking on disease processes and the effectiveness of treatment strategies. Summarization systems such as UpToDate provide manually curated overviews of clinical topics. However, given the expense associated with expert curation, utilizing natural language processing techniques for automatic summarization is an attractive alternative.

One approach to automatic summarization uses the relations found in the text to form summaries. Relation extraction is the process of automatically mining the input corpus for entities of interest (such as treatments and outcomes) and the semantic relationships that exist between them (such as "treatment X improves outcome Y"). Current relation extraction systems omit the context of the extracted relations. If a relation such as "treatment X improves outcome Y" is detected, this association is considered "true" regardless of the context in which the relation was found. However, context is crucial for capturing the full meaning of a relation.

Casama, a representation and summarization system for biomedical

* Corresponding author.
E-mail address: jigarcia@ucla.edu (J.I. Garcia-Gathright).

https://doi.org/10.1016/j.compbiomed.2017.10.034
Received 20 May 2017; Received in revised form 28 October 2017; Accepted 29 October 2017

literature on lung cancer, characterizes ”context” at two levels: the study level, which describes experimental conditions such as study design and outcome measures; and the patient/population level, which captures properties of the study population.

This paper describes an evaluation study that compared the summarization capabilities of Casama with a baseline system SemRep, a context-free representation. Manual and automatic annotations of several articles on driver mutations in cancer were reviewed and rated by multiple users. The results of the final analysis demonstrated significant advantages of Casama's contextualized relations over SemRep, particularly in the representation of strength of evidence.

2. Background and significance

2.1. Relation-based summarization

The representation of knowledge as concepts and relations was first explored in the 1970s by Novak, who applied this representation for education, and Sowa, who developed a computable formalism that supports querying and inference [1,2]. Relations (two or more concepts linked by a relationship to form a semantic unit) were proposed as the basic elements of knowledge. The collection of these relations, referred to as ”concept maps” or ”conceptual graphs” have been shown to be an effective way to represent, visualize, and communicate knowledge [3].

Many biomedical summarization systems use automatically extracted relations to structure their summaries. Some systems focus on mining biomedical articles for instances of a single relation type, such as protein-protein interactions [4–7], gene-protein interactions [8,9], drug-drug interactions [10–13], or treatment-disease relations [14,15]. Other summarization systems extract a variety of relations and present them visually to provide a comprehensive summary of the knowledge domain. For example, Telemakus uses relations extracted from tables and figures to represent claims in biomedical documents [16]. AliBaba uses pattern matching and co-occurrence filtering to extract protein-protein, gene--gene, and drug-disease relations, among others. These relations are then visualized as a graph for real-time browsing of PubMed query results [17]. BIOSQUASH, a summarizer based on the extraction of highly relevant sentences from the original document, produces a semantic graph to aid the sentence selection process [18]. Similarly, Morales et al. represent documents as a graph and cluster the sentences within the graph to determine which sentences are most significant [19].

The most significant work in visual summarization is the National Library of Medicine's Semantic MEDLINE. Semantic MEDLINE uses a relational framework based on SemRep to summarize claims made in scientific literature. Semantic MEDLINE utilizes four principles to select which relations or ”predications” should be included in the summary: relevance to the topic, connectivity of related predications, novelty of extracted knowledge, and salience or high frequency of predications within the source text. These are determined by examining the graph-based or statistical features of the semantic network [20].

Crucially, none of these systems use study context or patient/population context to focus their summaries. Indeed, while context in general has been explored in the domain of artificial intelligence [21–26], there has been relatively little development of context-sensitive systems to enhance biomedical relation extraction. Lussier et al. describe PhenoGO, a natural language processing system based on BioMedLEE, which assigns phenotypic context such as anatomical structure, body substance, and body system to Gene Ontology annotations [27]. Gerner et al. developed BioContext, a text mining system that contextualizes biomolecular events in terms of species involved, anatomical location, and speculation or negation [28]. BIOSMILE augments relations with the surrounding words signifying the location, manner, and timing of an event [29].

2.2. Evaluation of summarization systems

In a recent review of biomedical summarization systems, Mishra categorizes the evaluation of summarization systems into two groups: intrinsic and extrinsic [30]. Intrinsic methods assess the quality of summaries in terms of comprehensiveness, accuracy, and relevance with respect to a gold standard. As no reference standards exist for summarization in biomedicine, usually the gold standards used in evaluation are produced manually in a proprietary fashion. Alternatively, some systems use knowledge sources (such as the abstracts of papers) as their gold standard. Common evaluation metrics include precision and recall; in the case of text-based summaries, ROUGE metrics (Recall-Oriented Understudy for Gisting Evaluation) are often used [31]. Most of the systems reviewed by Mishra perform intrinsic evaluations. Extrinsic evaluations measure the task-oriented success of a system (e.g., time to completion, decision making accuracy, usability).

2.3. Significance

Casama builds upon current work in relation extraction by developing a framework in which the context of relations is represented and extracted, thus providing a more comprehensive summary that includes relevant knowledge such as experimental context and population attributes. The inclusion of additional knowledge in its summaries, and the tying of contextual knowledge to relations, can then be used to facilitate discovery of relevant facts by users.

Casama follows many of the summarization research trends identified by Mishra: aggregation of multiple documents to reveal current research directions, use of domain knowledge (i.e., Casama contexts) to enrich the summary semantically, and combination of lexical approaches with machine learning to extract relations and context. This paper presents an intrinsic evaluation of Casama's representation and its automatic extraction performance in terms of comprehensiveness and usability. This was accomplished by measuring user perceptions of summarization quality of manual and automatic annotations in comparison to a context-free representation, SemRep.

3. Materials and methods

3.1. Representations

3.1.1. Casama

During the initial design phase of the Casama representation, two lung cancer clinicians identified questions they perceived as important in a clinical study on driver mutations in cancer. These questions were: 1) how likely is it that my patient has this mutation; 2) is there a treatment available for this mutation; 3) is my patient likely to respond? Informed by these clinical questions, Casama was designed for the purpose of capturing knowledge related to four possible study objectives: mutation characterization (relevant to question 1), mutation detection (question 1), treatment (question 2), and prognosis (question 3). The Casama representation is composed of a set of relations that describe the main findings of a clinical study with respect to these objectives. Some examples of Casama relations are: *biomarker* **correlated with** *clinical feature*, *detection method* **detects** *biomarker*, *treatment* **improves** *outcome*, and *biomarker* **predicts** *outcome*.

Additionally, these relations are contextualized with patient context (i.e., attributes of the patient population such as biomarker status, disease stage, treatment history) and study context (e.g., methodological design, cohort size, endpoints measured). Contextualization enables these summaries to be queried from a patient-oriented and evidence-based perspective. For a detailed description of the Casama concepts, relations, and patient-oriented contexts, refer to [32]. Casama's representation of contexts related to strength of evidence can be found in Ref. [33].

3.1.2. SemRep

SemRep is a relation extraction system that parses biomedical text for subject-relation-object triples, which are presented in a context-free