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-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 manner. Its representation is based on relations found in the Unified Medical Language System Semantic Network, a comprehensive biomedical vocabulary [34,35]. SemRep was originally designed to facilitate biomedical research; however, it has been adapted for use in clinical applications. For example, Fiszman evaluated Semantic MEDLINE, a biomedical summarization system based on SemRep, in terms of ”clinical usefulness” in a drug-identification task [20,36]. A description of SemRep's schema of relations and concepts can be found in Ref. [37].

3.2. Evaluation

This evaluation was designed to measure and compare several aspects of each system from a user perspective: ability to represent the targeted knowledge domain comprehensively; effectiveness of the automatic annotation system; and overall usability.

The design for this user evaluation, depicted in Fig. 1, consists of the following steps. First, articles on a variety of topics were selected to form a gold standard. The articles cover topics outside the lung cancer domain to demonstrate Casama's generalizability. These articles were annotated both manually and automatically according to the Casama and SemRep representations. Then, users were recruited to review the articles and associated summaries. A questionnaire was composed that enables users to rate the summarization quality of Casama and SemRep on a number of topics, thus measuring the comprehensiveness of each system.

A statistical analysis was performed to discover whether one system rated significantly higher than the other. This analysis was performed both on individual articles and in aggregate. A separate analysis examined how summaries produced by manual annotation compared to those generated automatically.

3.2.1. Article collection and annotation

A variety of articles spanning multiple topics were selected for review. UpToDate and Medscape, two sources of human-curated summaries on a variety of clinical topics, were searched for articles on ”driver mutations in cancer” and ”targeted therapies in cancer.” The top ten relevant articles were selected to form the set of articles to be reviewed. The articles were numbered and organized into themes, as described in Table 1.

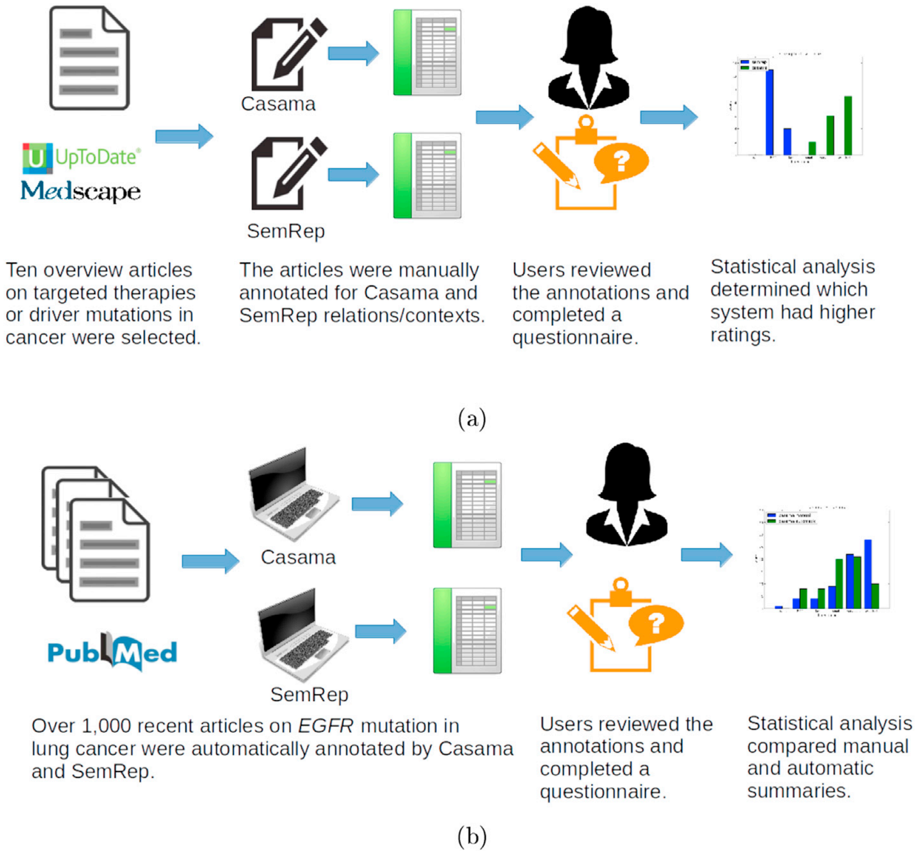
To discover how well the Casama and SemRep representations were able to capture the knowledge expressed in each article, the brat rapid annotation tool [38] was used to manually annotate each of these articles twice: first, for Casama relations and contexts; second, for SemRep relations. Annotation guidelines were used in both cases. The Casama guidelines for annotating relations and context can be found in Refs. [32,33]. Three sources were used for annotating SemRep relations: SemRep annotation guidelines detailed in Ref. [37], the existing SemRep gold standard [39], and the output of the SemRep relation extraction program.

The annotations were subsequently exported to spreadsheets containing each relation, the semantic types of its subject and object, the sentence in which the relation was found, and for Casama, the contexts in which the relation was found.

3.2.2. Automatic extraction

The most frequently occurring relations in the Casama gold standard were targeted for automatic extraction. The relations were: *biomarker* **correlated with** *clinical feature*, *detection method* **detects** *biomarker*, *biomarker* **predicts** *outcome*, *biomarker* **predicts worse** *outcome*, *biomarker* **does not predict** *outcome*, and *treatment* **associated with** *outcome*.

Similarly, a subset of clinical features/population context was selected based on an analysis of most frequently-occurring types in the Casama gold standard. This subset comprises: *stage*, *histology*, *biomarker*, and *treatment history*. *Sex*, *race*, and *smoking history* were also targeted as these were readily extracted by lexicon or regular expression. The most



Ten overview articles on targeted therapies or driver mutations in cancer were selected.

The articles were manually annotated for Casama and SemRep relations/contexts.

Users reviewed the annotations and completed a questionnaire.

Statistical analysis determined which system had higher ratings.

Over 1,000 recent articles on *EGFR* mutation in lung cancer were automatically annotated by Casama and SemRep.

Users reviewed the annotations and completed a questionnaire.

Statistical analysis compared manual and automatic summaries.

Fig. 1. Overview of the evaluation pipeline for (a) manually-annotated relations and (b) automatically-extracted relations.

Table 1
List of articles summarized and reviewed.

Article number	Title	Themes	Source
1	Anaplastic lymphoma kinase (*ALK*) fusion oncogene positive non-small cell lung cancer	Driver mutations in lung cancer	UpToDate
2	Systemic therapy for advanced non-small cell lung cancer with an activating mutation in the epidermal growth factor receptor	Driver mutations in lung cancer	UpToDate
3	Systemic treatment for *HER2*-positive metastatic breast cancer	Driver mutations in cancers other than lung	UpToDate
4	Anti-angiogenic and molecularly targeted therapy for advanced or metastatic clear-cell renal cell carcinoma	Driver mutations in cancers other than lung	UpToDate
5	Molecularly targeted therapy for metastatic melanoma	Driver mutations in cancers other than lung	UpToDate
6	Systemic therapy for the initial management of advanced non-small cell lung cancer without a driver mutation	Lung cancer, other topics	UpToDate
7	Advanced non-small cell lung cancer: Subsequent therapies for previously treated patients	Lung cancer, other topics	UpToDate
8	Personalized, genotype-directed therapy for advanced non-small cell lung cancer	Lung cancer, other topics	UpToDate
9	Genetics of Non-Small Cell Lung Cancer	Driver mutations in lung cancer	Medscape
10	Breast Cancer and *HER2*	Driver mutations in cancer	Medscape

common study design contexts (*study design*, *cohort size*, *p-values*, and *endpoints*) were extracted by a support vector machine classifier or by regular expression [40].

Automatic extraction of relations was performed by a combination of lexical matching (for concepts) and OpenIE, a general-domain relation extraction system [41]. A set of lexicons was developed from various sources such as biomedical ontologies [42–45], clinical guidelines [46], and papers [47,48]. These lexicons were used to tag instances of concepts within the Casama representation. In parallel, OpenIE extracted relations between noun phrases based on a lexico-syntactic parse of each sentence. If a relation was found whose concept types matched a frame in the Casama representation, it was tagged as a valid relation and included in the final summary. This process is illustrated diagrammatically in Fig. 2.

3.2.3. *Evaluation of automatically extracted relations*

To evaluate the summarization quality of automatically extracted relations and contexts, a document set consisting of recent articles on *EGFR* mutation in lung cancer was automatically annotated by both Casama and SemRep. The document set comprised 1340 articles from PubMed containing ”EGFR″ and ”lung” in the title or abstract, published between September 1, 2013 and December 15, 2015.

The automatic annotations were exported to the same spreadsheet format as described above.

3.2.4. *User assignments*

Seven users were recruited: three graduate students with 2–3 years of experience in medical informatics, and four researchers with 1–3 years of experience in a lung cancer clinic. Article assignments were allocated such that each article and its associated spreadsheets were viewed by 3–4 users.

To minimize variability among the users, three of the researchers evaluated the same set of articles (Group A); the three graduate students evaluated the remaining articles (Group B). The groups of articles were selected such that each group contained 1–2 articles from each article theme.

The seventh evaluator, a researcher, was assigned a set of higher difficulty articles and the spreadsheets of automatically extracted relations. ”Higher difficulty” articles included those in domains for which Casama was not tailored (i.e., ”driver mutations in non-lung cancers” and ”other topics in lung cancer”). As a result, we maximized the number of users viewing these more challenging articles.

The final user assignments were as follows. Group A evaluated the summaries for articles 1, 3, 4, 6, and 9. Group B reviewed articles 2, 5, 7, 8, 10, and the automatic annotations. The seventh evaluator was assigned articles 2, 3, 4, 7, and the automatic annotations.

3.2.5. *Questionnaire*

A user questionnaire was designed to measure the quality of the SemRep and Casama relations. A variety of topics were chosen to evaluate the comprehensiveness of each representation over the target domain. We examined the SemRep and Casama concepts, relations, and

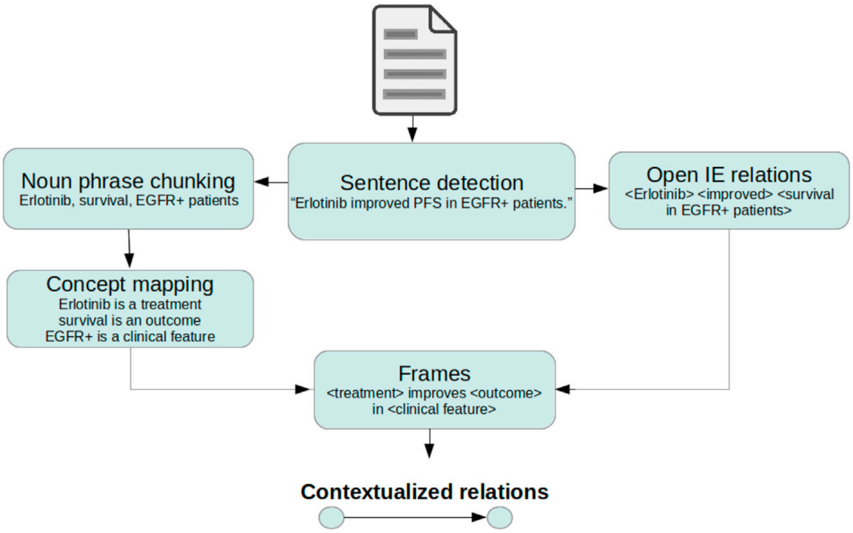

Noun phrase chunking
Erlotinib, survival, EGFR+ patients
Sentence detection
"Erlotinib improved PFS in EGFR+ patients."
Open IE relations
<Erlotinib> <improved> <survival in EGFR+ patients>
Concept mapping
Erlotinib is a treatment
survival is an outcome
EGFR+ is a clinical feature
Frames
<treatment> improves <outcome> in <clinical feature>
Contextualized relations

Fig. 2. Overview of the automatic extraction method. First, abstracts are pre-processed to extract sentences and noun phrases. Lexicons are used to identify concepts from the noun phrase chunks. In parallel, OpenIE is used to extract relations from raw sentences in a domain-independent manner. Tagged concepts and relations are mapped to frames to produce the final relations.

contexts found in the articles to identify the types of knowledge they attempt to represent. We then designed questions that would determine the success of each system at capturing that knowledge. The relations and contexts covered were translated into specific topics to be evaluated by users: identification of drugs/treatments, effectiveness of drugs/treatments, clinical guidelines for drugs/treatments, side effects of drugs/treatments, identification of genes/biomarkers, prognostic effects of genes/biomarkers, clinical characteristics of genes/biomarkers, biochemical characteristics of genes/biomarkers, diagnostic tests/detection methods, and strength of evidence. These topics represent the union of knowledge captured over both systems; thus, they range from biological topics (such as gene identification) to clinical topics (such as prognosis). We aimed to discover the overlap between systems, identify the relative advantages of each system, and compare the success of each system from a user-oriented perspective. The users rated the quality of the SemRep relations and Casama contextualized relations with respect to these topics on a 5-point Likert scale (5 = excellent, 4 = very good, 3 = good, 2 = fair, 1 = poor).

Additionally, the users rated the overall summarization quality, comprehensibility, and usefulness of SemRep and Casama for several high-level applications: clinical decision support, precision medicine, evidence based medicine, meta-analysis, and general biomedical research.

In a free-text portion of the questionnaire, users were asked to state what relevant information was missing from the SemRep and Casama summaries. Additional free-text comments on any topic relating to SemRep and Casama were also encouraged. These comments were then interpreted to provide insights linking the quantitative measurements to the qualitative merits of each representation, and to inform future work.

The full text of the questionnaire can be found in the Appendix.

3.2.6. Analysis

The Wilcoxon rank sums test, a non-parametric test commonly used to analyze ordinal data, determined whether one representation tended to have higher scores than another. A significance threshold of 0.05 was selected; however, due to the large number of hypothesis tests (one for every topic and article), Bonferroni correction was applied, resulting in various p-value thresholds for each group of tests [49].

To achieve sufficient statistical power for each topic, responses were aggregated over all articles (p-value threshold = 0.05/17 = 0.003). To assess the quality of automatic summarization with respect to ”ideal” summaries produced by manual annotation, the ratings for automatic extraction of relations by SemRep and Casama were compared to that of manual annotation (p-value threshold = 0.05/17 = 0.003).

4. Results

4.1. SemRep vs. Casama per topic over all articles

Tables 2 and 3 present the median scores for each topic when aggregated over all articles and the p-values for the Wilcoxon rank sums test. In absolute terms, SemRep received its highest scores for identification of drugs/treatments and identification of genes/biomarkers. Users rated Casama most highly for identification of drugs/treatments, clinical guidelines, prognostic effects of genes, and strength of evidence.

Comparing performance between Casama and SemRep, Groups A and B rated Casama significantly more positively in effectiveness of drugs/treatments, clinical guidelines, prognostic effects of genes/biomarkers, and strength of evidence. Casama also received higher scores for overall summarization quality, comprehensibility, usability, and in all high-level applications.

4.2. Automatic extraction

Table 4 presents the p-values for the Wilcoxon rank sums test comparing SemRep's automatically extracted relations to Casama's automatically extracted relations and contexts. No individual topic showed significantly higher Casama scores; however, Casama did outperform SemRep when aggregated over all topics ($p = 4.7$ e-05).

Table 2
Group A: Median scores and p-values for Wilcoxon rank sums when aggregated over all articles, testing whether Casama's scores tend to be higher than SemRep's. Bold: $p \leq 0.003$ (significant after Bonferroni correction).

Topic	Median (SemRep)	Median (Casama)	p-value
Identification of drugs	4	4	0.079
Effectiveness of drugs	2	4	**2.8e-06**
Clinical guidelines	2	4	**6.5e-06**
Side effects	3	4	0.16
Identification of genes	3	3	0.044
Prognostic effects of genes	2	3	**0.00032**
Clinical characteristics of genes	2	3	0.027
Biochemical characteristics of genes	2	3	0.22
Diagnostic tests	2	3	0.55
Strength of evidence	2	4	**6.5e-07**
Overall summarization quality	2	4	**1.5e-05**
Comprehensibility	2	4	**0.00089**
Clinical decision support	2	4	**4.3e-06**
Precision medicine	2	3	**1.8e-05**
Evidence-based medicine	2	4	**1.7e-05**
Meta-analysis	2	4	**3.6e-05**
General biomedical research	2	4	**0.00032**

Table 3
Group B: Median scores and p-values for Wilcoxon rank sums when aggregated over all articles, testing whether Casama's scores tend to be higher than SemRep's. Bold: $p \leq 0.003$ (significant after Bonferroni correction).

Topic	Median (SemRep)	Median (Casama)	p-value
Identification of drugs	4	5	0.026
Effectiveness of drugs	2	4	**7.5e-07**
Clinical guidelines	2	5	**2.2e-06**
Side effects	2	3	0.014
Identification of genes	4	4	0.26
Prognostic effects of genes	1	5	**4.8e-05**
Clinical characteristics of genes	1	3	0.013
Biochemical characteristics of genes	1	1	0.89
Diagnostic tests	1	3	0.11
Strength of evidence	1	5	**6.5e-07**
Overall summarization quality	2	4	**1.1e-06**
Comprehensibility	2	3.5	**7.6e-05**
Clinical decision support	2	5	**6.0e-05**
Precision medicine	2	4	**4.8e-05**
Evidence-based medicine	2	5	**1.4e-06**
Meta-analysis	2	4	**2.5e-05**
General biomedical research	2	4	**8.0e-05**

Finally, manual annotations were compared with automatically extracted relations for both SemRep and Casama (Table 5). SemRep's automatic extraction of relations on *EGFR* mutation in lung cancer was not significantly different from manual annotations on the same topic, either individually or in aggregate.

Similarly for Casama, no significant effects were seen for the individual topics. However, when aggregating Casama scores over all topics, manual annotation of Casama relations/contexts significantly

Table 4
P-values for Wilcoxon rank sums, testing whether Casama's scores for automatically extracted relations and contexts tend to be higher than SemRep's. Bold: $p \leq 0.003$ (significant after Bonferroni correction).

Topic	p-value
Identification of drugs	0.89
Effectiveness of drugs	0.043
Clinical guidelines	0.56
Side effects	0.15
Identification of genes	0.19
Prognostic effects of genes	0.083
Clinical characteristics of genes	0.043
Biochemical characteristics of genes	0.89
Diagnostic tests	0.89
Strength of evidence	0.11
Overall summarization quality	0.043
Comprehensibility	0.08
Clinical decision support	0.15
Precision medicine	0.06
Evidence-based medicine	0.25
Meta-analysis	0.06
General biomedical research	0.47
All topics	**4.7e-05**

Table 5
P-values for Wilcoxon rank sums, comparing manual annotation to automatic extraction. Bold: $p \leq 0.003$ (significant after Bonferroni correction).

Topic	Manual vs. automatic (SemRep)	Manual vs. automatic (Casama)
Identification of drugs	0.89	0.25
Effectiveness of drugs	1.0	0.56
Clinical guidelines	0.25	0.19
Side effects	0.89	0.03
Identification of genes	0.56	0.56
Prognostic effects of genes	0.67	0.31
Clinical characteristics of genes	0.56	0.25
Biochemical characteristics of genes	1.0	0.89
Diagnostic tests	0.021	0.31
Strength of evidence	1.0	0.25
Overall summarization quality	1.0	0.15
Comprehensibility	0.22	0.38
Clinical decision support	0.77	0.083
Precision medicine	1.0	0.39
Evidence-based medicine	0.89	0.39
Meta-analysis	0.67	0.39
General biomedical research	0.89	0.31
All topics	0.67	**0.00061**

outperformed automatic extraction ($p = 0.00061$).

4.3. *Free-text comments*

The free-text comments generally favored Casama over SemRep, which is consistent with higher ratings on overall quality and comprehensibility. The users' comments give insight as to why: four users noted a larger number of concept and relation types with SemRep, resulting in more difficulty finding the desired information. In contrast, users described Casama's representation as ”concise” and ”easy to search.” Three users felt that the relations targeted by SemRep (such as ”chemotherapy treats human patients”) were ”very broad,” whereas Casama was described as ”detailed” and ”focused.” Two users stated that the SemRep relations were ”repetitive,” which one user wrote might be useful for users unfamiliar with the knowledge space.

Certain Casama concept and relation types were identified as being most useful: *biomarkers*, *treatments*, and *outcomes* were called out specifically for helping users locate relevant information. In particular, users appreciated that each concept type was unique (as opposed to SemRep, in which multiple semantic types such as pharmacologic substance and therapeutic procedure both refer to treatment). Users also named the **improves** and **recommended for** relations as very helpful in capturing effectiveness of drugs and associated clinical guidelines; these comments are consistent with higher ratings in the questionnaire on these topics.

Casama's contextual elements were also viewed positively by the users, especially contexts related to strength of evidence. Nearly every user expressed appreciation for ”clinical trial information” or ”whether or not a clinical trial determined the results.” One user said of Casama's representation for strength of evidence, ”This is where Casama excels.” There were fewer comments related to Casama's representation of patient characteristics, though these were generally favorable as well (”stage of cancer or past treatment is very helpful,” ”Casama provided more nuance in clinical characteristics of genes/biomarkers.”) In contrast, users noted that SemRep does not capture these contexts; therefore, users provided lower ratings on these topics.

Lastly, users gave examples of relevant information that was not targeted by SemRep or Casama. These included: statistical details (p-value thresholds, hazard ratios), knowledge spanning multiple sentences, names of clinical trials, names of agencies with published guidelines, and further information about diagnostic tests.

5. Discussion

This user evaluation showed that Casama's contextualized relations resulted in more targeted and usable summaries, as evidenced in both the user ratings and the free-text comments. Casama did better than SemRep in the representation of clinical information such as treatment effectiveness, prognosis, and especially strength of evidence, which was highly rated both quantitatively and qualitatively. Indeed, Casama was designed specifically to address these clinical information needs through its concepts, relations, and contexts; this evaluation shows that it succeeds in doing so. In contrast, SemRep's relations are based on a broad biomedical vocabulary. This is consistent with positive performance by SemRep in the representation of general biomedical knowledge (e.g., identification of drugs/treatments and genes/biomarkers). Notably, SemRep did not outperform Casama in these cases, suggesting that in the domain of driver mutations in cancer, Casama captures the same knowledge as SemRep and more.

There were also topics for which neither system performed particularly well. These include: side effects, clinical characteristics of genes/biomarkers, biochemical characteristics of genes/biomarkers (targeted by SemRep but not Casama), and diagnostic tests. For these topics, neither SemRep nor Casama received high marks, nor did one system outperform the other. Iterating on the representation of these topics based on user opinion, expert knowledge, and examination of existing data and ontologies would likely improve Casama's performance on these topics. For example, Casama's representation does not represent side effects explicitly, instead combining the *outcome* concept to include both positive and negative outcomes. A more nuanced representation (perhaps based on the Ontology of Adverse Events [50] or bootstrapped from the Federal Drug Administration's Adverse Events Reporting System [51]) may be valuable to users seeking this type of information. For diagnostic tests, users provided suggestions in the free-text comments, noting that scoring mechanisms and clinical recommendations were not captured. Clinical guidelines for diagnosis [52] could help inform improvements of the Casama representation on this topic. In terms of clinical characteristics, Casama does include a fairly granular representation, though some areas could be improved. For example, one user noted in the free-text comments that patient age was not clearly specified (e.g., ”younger” vs. ”older” patients was not defined). An expert-guided review could provide further recommendations for improving Casama's representation in this area.

The free-text comments also revealed some useful observations about what users value in a summarization system in terms of usability. Users appreciated that Casama provided a small number of specific concept types, as this improved searchability of the summaries. In contrast, SemRep's concept types proved too numerous for purposes of summarization. Similarly, the Casama representation contains fewer and more targeted relation types, whereas SemRep's summaries include what one user referred to as ”basic facts” such as ”non small-cell lung cancer is a neoplastic process.” This suggests that in the context of a particular task, a coarse but targeted representation outweighs granularity from a user perspective; this observation can be used when designing representations for summarization in general.

With respect to automatic annotation, an important observation is that SemRep's manual annotations are not significantly different from SemRep's automatic annotations; in contrast, Casama's manual annotations outperformed Casama's automatically extracted relations. Therefore, SemRep is more limited by its representation than its relation extraction method; the opposite is true for Casama. Another priority for future work would be improvement of Casama's relation extraction algorithms. Despite this, Casama's summaries produced via automatic extraction outperformed those of SemRep.

This evaluation showed that Casama's contextualized approach to summarization adds substantial value to summarization applications, particularly for clinical decision support systems seeking to facilitate evidence-based and precision medicine. Strength of evidence and population contexts are not only useful, but necessary elements in clinical decision support systems that examine biomedical literature. Beyond the clinical domain, Casama's contextualized representation would be useful to users such as researchers, educators, and consumers, whose health-related searches could be targeted more meaningfully with contextual knowledge on study populations or strength of evidence.

5.1. Related work

As is most common with the evaluation of summarization systems [30], we performed an intrinsic evaluation [53]. Unlike other intrinsic evaluations that measured summarization quality via precision/recall, ROUGE metrics, or custom metrics [36,54,55], our study elicited quality ratings from users of Casama and SemRep. Our evaluation was similar to previous evaluations of clinical summarization systems in that participants manually rated the quality of summarized information using a discrete scale. For example, Kushniruk et al. and Elhadad et al. performed user studies to evaluate usability of a clinical/consumer health summarization system [56,57]; in Ref. [58], physicians rated a clinical question answering system in terms of usability, answer quality, and time spent. A differentiating feature of our evaluation is that we compared our representation (Casama) against an existing benchmark representation (SemRep) with two different types of relations: (i) relations that were manually annotated by humans and (ii) relations that were automatically extracted by a machine. This evaluation design thus helped to pinpoint which of Casama's strengths are due specifically to the design of its schema.

5.2. Limitations and future work

A limitation of this study design is the relatively small number of evaluators and the finite amount of time available to review the articles and summaries. Consequently, no article was reviewed by all evaluators, potentially resulting in variability between Groups A and B. While ratings were consistent between Groups A and B, a larger evaluation with more reviewers per article would be ideal. An evaluation with clinical experts would also reveal further insights about the summarization systems.

While the authors attempted to be as objective as possible in creating the annotations and designing the questionnaire, it is impossible to be perfectly unbiased. Furthermore, each evaluator was previously familiar with the SemRep or Casama representations, making it impossible to perform a blinded study. To mitigate the effects of inherent bias causing the experimental design to favor one system over another, a future evaluation should utilize external, blinded annotators and a questionnaire based on an expert-curated set of summarization requirements.

Short-term improvements to Casama include the incorporation of the new concept types identified by the evaluators (statistical concepts, additional clinical trial metadata, information about adverse events, and details of diagnostic tests), and iteration on the automatic extraction algorithm. With respect to evaluation, the promising results of this study suggest that an extrinsic evaluation of Casama in the clinical setting is warranted, as this would demonstrate how the system affects the performance of clinical tasks, as well as how the system could be integrated into existing clinical workflows. An example would be to evaluate the speed with which a physician identifies the optimal treatment for lung cancer patients, given each patient's particular medical history and genetic profile. Because Casama uses a graph-based representation, it is particularly amenable to visualization, and thus to assessing the impact of such visualizations on clinical tasks – an area that has been identified as needing further research [30].

6. Conclusion

This paper described an evaluation study in which users rated the summarization quality of Casama and SemRep. While both representations achieved high scores for identification of drugs and genes, Casama outperformed SemRep in capturing knowledge related to strength of evidence, drug effectiveness, clinical guidelines, and more. Casama was also highly rated for overall summarization quality and applications such as evidence-based medicine. Further development of Casama's automatic extraction algorithms and a clinically-oriented evaluation are warranted.

Acknowledgements

This work was supported by NLM T15-LM007356, NIH/NLM R01-LM009961, NIH T32-EB016640-02, NIH CA208403, and the UCLA Department of Radiological Sciences.

APPENDIX

Questionnaire.

Name of article:
Rating labels: 5 = excellent, 4 = very good, 3 = good, 2 = fair, 1 = poor.
Query/interaction task
Rate the quality of the relations for information on…
Identification of drugs/treatments (e.g., which drugs/treatments are used in this domain?)
Effectiveness of drugs/treatments (e.g., how well do these drugs/treatments work?)
Clinical guidelines for drugs/treatments (e.g., in what situations should these drugs/treatments be administered?)
Side effects of drugs/treatments (e.g., what side effects are associated with these drugs/treatments?)

Identification of genes/biomarkers (e.g., which genes/biomarkers are relevant in this domain?)
Prognostic effects of genes/biomarkers (e.g., which genes/biomarkers predict response to treatment?)
Clinical characteristics of genes/biomarkers (e.g., who is more likely to have this gene/biomarker?)
Biological characteristics of genes/biomarkers (e.g., which biomolecular pathways are affected by this gene/biomarker?)
Diagnostic tests/detection methods (e.g., which procedures are used to detect this gene/biomarker?)
Strength of evidence (e.g., which results have been demonstrated in large clinical trials?)
Overall impressions
Rate the overall quality of the relations for capturing the knowledge contained in the article.
Rate the comprehensibility of the relations.
Rate the usefulness of the relations for informing clinical decision support.
Rate the usefulness of the relations for informing precision medicine.
Rate the usefulness of the relations for informing evidence based medicine.
Rate the usefulness of the relations for informing meta-analysis.
Rate the usefulness of the relations for informing general biomedical research.
Overall comments:

References

[1] Joseph Donald Novak, A Theory of Education, 1977.

[2] John F. Sowa, Conceptual graphs for a data base interface, IBM J. Res. Dev. 20 (4) (1976) 336–357.

[3] Joseph D. Novak, D. Bob Gowin, Learning How to Learn, Cambridge University Press, 1984.

[4] Christian Blaschke, Miguel A. Andrade, Christos A. Ouzounis, Alfonso Valencia, Automatic extraction of biological information from scientific text: protein-protein interactions, Ismb 7 (1999) 60–67.

[5] Toshihide Ono, Haretsugu Hishigaki, Akira Tanigami, Toshihisa Takagi, Automated extraction of information on protein–protein interactions from the biological literature, Bioinformatics 17 (2) (2001) 155–161.

[6] Raymond J. Mooney, Razvan C. Bunescu, Subsequence kernels for relation extraction, Adv. neural Inf. Process. Syst. (2005) 171–178.

[7] Lei Hua, Chanqin Quan, A shortest dependency path based convolutional neural network for protein-protein relation extraction, BioMed Res. Int. 2016 (2016).

[8] Katrin Fundel, Robert Küffner, Ralf Zimmer, Relex – relation extraction using dependency parse trees, Bioinformatics 23 (3) (2006) 365–371.

[9] Claudio Giuliano, Alberto Lavelli, Lorenza Romano, Exploiting shallow linguistic information for relation extraction from biomedical literature, EACL 18 (2006) 401–408. Citeseer.

[10] Isabel Segura-Bedmar, Paloma Martinez, Cesar de Pablo-Sánchez, Using a shallow linguistic kernel for drug–drug interaction extraction, J. Biomed. Inf. 44 (5) (2011) 789–804.

[11] Zibo Yi, Shasha Li, Jie Yu, Qingbo Wu, Drug-drug Interaction Extraction via Recurrent Neural Network with Multiple Attention Layers, arXiv preprint arXiv: 1705.03261, 2017.

[12] Sunil Kumar Sahu, Ashish Anand, Drug-drug Interaction Extraction from Biomedical Text Using Long Short Term Memory Network, arXiv preprint arXiv: 1701.08303, 2017.

[13] Zhehuan Zhao, Zhihao Yang, Ling Luo, Hongfei Lin, Jian Wang, Drug drug interaction extraction from biomedical literature using syntax convolutional neural network, Bioinformatics 32 (22) (2016) 3444–3453.

[14] Barbara Rosario, Marti A. Hearst, Classifying semantic relations in bioscience texts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2004, p. 430.

[15] Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, Hans-Peter Kriegel, Extraction of semantic biomedical relations from text using conditional random fields, BMC Bioinforma. 9 (1) (2008) 207.

[16] Sherrilynne S. Fuller, Debra Revere, Paul F. Bugni, George M. Martin, A knowledgebase system to enhance scientific discovery: Telemakus, Biomed. Digit. Libr. 1 (1) (2004) 2.

[17] Conrad Plake, Torsten Schiemann, Marcus Pankalla, Jörg Hakenberg, Ulf Leser, Alibaba: pubmed as a graph, Bioinformatics 22 (19) (2006) 2444–2445.

[18] Zhongmin Shi, Gabor Melli, Yang Wang, Yudong Liu, Baohua Gu, Mehdi M. Kashani, Anoop Sarkar, Fred Popowich, Question answering summarization of multiple biomedical documents, in: Advances in Artificial Intelligence, Springer, 2007, pp. 284–295.

[19] Laura Plaza Morales, Alberto Díaz Esteban, Pablo Gervás, Concept-graph based biomedical automatic summarization using ontologies, in: Proceedings of the 3rd Textgraphs Workshop on Graph-based Algorithms for Natural Language Processing, Association for Computational Linguistics, 2008, pp. 53–56.

[20] Thomas C. Rindflesch, Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat, Dongwook Shin, H. Kilicoglu, M. Fiszman, G. Rosemblat, D. Shin, Semantic medline: an advanced information management application for biomedicine, Inf. Serv. Use 31 (1–2) (2011) 15–21.

[21] John McCarthy, Notes on Formalizing Context, 1993.

[22] Douglas B. Lenat, Cyc: a large-scale investment in knowledge infrastructure, Commun. ACM 38 (11) (1995) 33–38.

[23] Eckart Walther, Henrik Eriksson, Mark A. Musen, Plug-and-Play: construction of task-specific expert-system shells using sharable context ontologies, in: Proceedings of the AAAI Workshop on Knowledge Representation Aspects of Knowledge Acquisition, Citeseer, 1992, pp. 191–198.

[24] Peter Turney, The Identification of Context-Sensitive Features: a Formal Definition of Context for Concept Learning, 1996.

[25] Fausto Giunchiglia, Contextual reasoning, Epistemologia, special issue I Linguaggi e le Macchine 16 (1993) 345–364.

[26] Patrick Brézillon, Focusing on context in human-centered computing, IEEE Intell. Syst. 18 (3) (2003) 62–66.

[27] Lee T. Sam, Eneida A. Mendonça, Jianrong Li, Judith Blake, Carol Friedman, Yves A. Lussier, Phenogo: an integrated resource for the multiscale mining of clinical and biological data, Bmc Bioinforma. 10 (Suppl 2) (2009) S8.

[28] Martin Gerner, Farzaneh Sarafraz, Casey M. Bergman, Goran Nenadic, Biocontext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events, Bioinformatics 28 (16) (2012) 2154–2161.

[29] Richard TH. Tsai, Wen-Chi Chou, Ying-Shan Su, Yu-Chun Lin, Cheng-Lung Sung, Hong-Jie Dai, Irene TH. Yeh, Wei Ku, Ting-Yi Sung, Wen-Lian Hsu, Biosmile: a semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features, BMC Bioinforma. 8 (1) (2007) 325.

[30] Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R. Weir, Siddhartha Jonnalagadda, Javed Mostafa, Guilherme Del Fiol, Text summarization in the biomedical domain: a systematic review of recent research, J. Biomed. Inf. 52 (2014) 457–467.

[31] Chin-Yew Lin, Rouge: a package for automatic evaluation of summaries, Text Summ. branches out Proc. ACL-04 workshop 8 (2004).

[32] Jean I. Garcia-Gathright, Nicholas J. Matiasz, Edward B. Garon, Denise R. Aberle, Ricky K. Taira, Alex AT. Bui, Toward patient-tailored summarization of lung cancer literature, in: In Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on, IEEE, 2016, pp. 449–452.

[33] Jean I. Garcia-Gathright, Andrea Oh, Phillip A. Abarca, Mary Han, William Sago, Marshall L. Spiegel, Brian Wolf, Edward B. Garon, Alex AT. Bui, Denise R. Aberle, Representing and extracting lung cancer study metadata: study objective and study design, Comput. Biol. Med. 58 (2015) 63–72.

[34] Thomas C. Rindflesch, Marcelo Fiszman, Bisharah Libbus, Semantic interpretation for the biomedical research literature, in: Medical Informatics, Springer, 2005, pp. 399–422.

[35] Olivier Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids Res. 32 (suppl 1) (2004) D267–D270.

[36] Marcelo Fiszman, Dina Demner-Fushman, Halil Kilicoglu, Thomas C. Rindflesch, Automatic summarization of medline citations for evidence-based medical treatment: a topic-oriented evaluation, J. Biomed. Inf. 42 (5) (2009) 801–813.

[37] Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, Thomas C. Rindflesch, Constructing a semantic predication gold standard from the biomedical literature, BMC Bioinforma. 12 (1) (2011) 486.

[38] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, Jun'ichi Tsujii, Brat: a web-based tool for nlp-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2012, pp. 102–107.

[39] Lister Hill National Center for Biomedical Communications, Semantic Knowledge Representation, 2013. Accessed: 06 june 2016, https://semrep.nlm.nih.gov/GoldStandard.html.

[40] Thorsten Joachims, Learning to Classify Text Using Support Vector Machines, Springer, Boston, 2002 edition edition, April 2002.

[41] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, Oren Etzioni, Open information extraction for the web, IJCAI 7 (2007) 2670–2676.

[42] Anna L. Rich, Laila J. Tata, Rosamund A. Stanley, Catherine M. Free, Michael D. Peake, David R. Baldwin, Richard B. Hubbard, Lung cancer in england: information from the national lung cancer audit (lucada), Lung cancer 72 (1) (2011) 16–22.

[43] Kevin Donnelly, Snomed-ct: the advanced terminology and coding system for ehealth, Stud. health Technol. Inf. 121 (279) (2006).

[44] William D. Travis, Elisabeth Brambilla, Masayuki Noguchi, Andrew G. Nicholson, Kim Geisinger, Yasushi Yatabe, Charles A. Powell, David Beer, Greg Riely, Kavita Garg, et al., International association for the study of lung cancer/american thoracic society/european respiratory society: international multidisciplinary classification of lung adenocarcinoma: executive summary, Proc. Am. Thorac. Soc. 8 (5) (2011) 381–385.

[45] Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, Lawrence W. Wright, Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information, J. Biomed. Inf. 40 (1) (2007) 30–43.

[46] National Cancer Institute, Levels of Evidence: Adult and Pediatric Treatment Studies, 2015. Accessed: 18 September 2015, http://www.cancer.gov/publications/pdq/levels-evidence/treatment.

[47] Thomas A. Hensing, Ravi Salgia, Molecular biomarkers for future screening of lung cancer, J. Surg. Oncol. 108 (5) (2013) 327–333.

[48] Gillian Ellison, Guanshan Zhu, Alexandros Moulis, Simon Dearden, Georgina Speake, Rose McCormack, Egfr mutation testing in lung cancer: a review of available methods and their use for analysis of tumour tissue and cytology samples, J. Clin. Pathol. 66 (2) (2013) 79–89.

[49] John H. McDonald, Handbook of Biological Statistics, vol. 2, Sparky House Publishing Baltimore, MD, 2009.

[50] Yongqun He, Sirarat Sarntivijai, Yu Lin, Zuoshuang Xiang, Abra Guo, Shelley Zhang, Desikan Jagannathan, Luca Toldo, Cui Tao, Barry Smith, Oae: the ontology of adverse events, J. Biomed. Semant. 5 (1) (2014) 29.

[51] Toshiyuki Sakaeda, Akiko Tamon, Kaori Kadoyama, Yasushi Okuno, Data mining of the public version of the fda adverse event reporting system, Int. J. Med. Sci. 10 (7) (2013) 796.

[52] S. Novello, F. Barlesi, Raffaele Califano, T. Cufer, S. Ekman, M. Giaj Levra, K. Kerr, S. Popat, M. Reck, S. Senan, et al., Metastatic non-small-cell lung cancer: esmo clinical practice guidelines for diagnosis, treatment and follow-up, Ann. Oncol. 27 (suppl_5) (2016) v1–v27.

[53] Karen Sparck Jones, Julia R. Galliers, Evaluating Natural Language Processing Systems: an Analysis and Review, vol. 1083, Springer Science & Business Media, 1995.

[54] Ping Chen, Rakesh Verma, A query-based medical information summarization system using ontology knowledge, in: In Computer-based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on, IEEE, 2006, pp. 37–42.

[55] Diego Mollá, María Elena Santiago-Martínez, Creation of a corpus for evidence based medicine summarisation, Australas. Med. J. 5 (9) (September 2012) 503–506.

[56] Andre W. Kushniruk, Min-Yem Kan, Kathleen McKeown, Judith Klavans, Desmond Jordan, Mark LaFlamme, Vimia L. Patel, Usability evaluation of an experimental text summarization system and three search engines: implications for the reengineering of health care interfaces, in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2002, p. 420.

[57] Noemie Elhadad, M.-Y. Kan, Judith L. Klavans, K.R. McKeown, Customization in a unified framework for summarizing medical literature, Artif. Intell. Med. 33 (2) (2005) 179–198.

[58] Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J. Cimino, John Ely, Hong Yu, Askhermes: an online question answering system for complex clinical questions, J. Biomed. Inf. 44 (2) (2011) 277–288.