Lossio-Ventura *et al. BMC Medical Informatics and Decision Making* 2018, **18**(Suppl 2):55
https://doi.org/10.1186/s12911-018-0635-5

BMC Medical Informatics and Decision Making

RESEARCH Open Access

OC-2-KB: integrating crowdsourcing into an obesity and cancer knowledge base curation system

Juan Antonio Lossio-Ventura[1], William Hogan[1], François Modave[1], Yi Guo[1], Zhe He[2], Xi Yang[1], Hansi Zhang[1] and Jiang Bian[1]*

From The 2nd International Workshop on Semantics-Powered Data Analytics
Kansas City, MO, USA. 13 November 2017

Abstract

Background: There is strong scientific evidence linking obesity and overweight to the risk of various cancers and to cancer survivorship. Nevertheless, the existing online information about the relationship between obesity and cancer is poorly organized, not evidenced-based, of poor quality, and confusing to health information consumers. A formal knowledge representation such as a Semantic Web knowledge base (KB) can help better organize and deliver quality health information. We previously presented the OC-2-KB (Obesity and Cancer to Knowledge Base), a software pipeline that can automatically build an obesity and cancer KB from scientific literature. In this work, we investigated crowdsourcing strategies to increase the number of ground truth annotations and improve the quality of the KB.

Methods: We developed a new release of the OC-2-KB system addressing key challenges in automatic KB construction. OC-2-KB automatically extracts semantic triples in the form of subject-predicate-object expressions from PubMed abstracts related to the obesity and cancer literature. The accuracy of the facts extracted from scientific literature heavily relies on both the quantity and quality of the available ground truth triples. Thus, we incorporated a crowdsourcing process to improve the quality of the KB.

Results: We conducted two rounds of crowdsourcing experiments using a new corpus with 82 obesity and cancer-related PubMed abstracts. We demonstrated that crowdsourcing is indeed a low-cost mechanism to collect labeled data from non-expert laypeople. Even though individual layperson might not offer reliable answers, the collective wisdom of the crowd is comparable to expert opinions. We also retrained the relation detection machine learning models in OC-2-KB using the crowd annotated data and evaluated the content of the curated KB with a set of competency questions. Our evaluation showed improved performance of the underlying relation detection model in comparison to the baseline OC-2-KB.

Conclusions: We presented a new version of OC-2-KB, a system that automatically builds an evidence-based obesity and cancer KB from scientific literature. Our KB construction framework integrated automatic information extraction with crowdsourcing techniques to verify the extracted knowledge. Our ultimate goal is a paradigm shift in how the general public access, read, digest, and use online health information.

Keywords: Semantic web knowledge base, Information extraction, Biomedical named-entity recognition, Relation extraction, Crowdsourcing, Obesity, Cancer

*Correspondence: bianjiang@ufl.edu
[1]Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, 2004 Mowry Road, 32610 Gainesville, FL, USA
Full list of author information is available at the end of the article

Background

Overweight and obesity are associated with 2.8 million deaths throughout the world. More than 1.9 billion adults (39% of adults) were overweight in 2014; of which, over 600 million (13% of adults) were obese. More than one-third (34.9% or 78.6 million) of US adults are obese and the prevalence of obesity among children is also increasing [1]. As a result of many meta- and pooled analyses on obesity-related cancers, existing knowledge and data on obesity and the increased risk it poses for cancer, are proliferating. Obesity and overweight are related with major risk factors for cancers of the endometrium, breast, kidney, colorectal, pancreas, esophagus, ovaries, gallbladder, thyroid, and possibly prostate [2–5]. Obesity accounts for 3–10% of cancer cases and deaths [6, 7], and it will be linked to more cancer cases than smoking within ten years [4].

On the other hand, interventions that may effectively modify excess weight can be used for cancer control, preventing further cancer burden [8]. Many health behavior theories, such as the information-motivation-behavioral skills mode (IMB) [9] and the integrated behavior model (IBM) [10], recognized that an individual needs the information and knowledge to initiate and perform health behavior changes towards healthier lifestyles. Increasingly, people engage in health information seeking via the Internet. In the US, 87% of adults have Internet access and 72% look online for health information [11]. Meanwhile, the quality of online health information varies widely [12–14]. In particular, existing online information on obesity, especially its relationship to cancer, is heterogeneous ranging from pre-clinical models to case studies to mere hypothesis-based arguments. While the direct causal relationship between obesity and cancer has been difficult to definitively prove, research studies have generated a tremendous amount of supporting data. But collectively, these data are poorly organized in the public domain. Typical consumers cannot translate the vast amounts of online health information into usable knowledge nor can they assess its quality. Online health information consumers face a number of access barriers including information overload and disorganization, terminology and language barriers, lack of user friendliness, and inconsistent updates [13, 15, 16].

There is an urgent need to organize the vast and increasing amount of information on possible links between obesity and cancer in a way that helps consumers understand and use the information in a meaningful way. This involves collating the information, linking it to evidence in the scientific literature, evaluating its quality, and presenting high quality information relevant to their specific questions in a user-friendly way. Thus, we seek to fill critical gaps in knowledge representation of obesity and its relationship to cancer, improve dissemination of knowledge in obesity and cancer research, and ultimately to provide the general public with a knowledge base (KB) of well-organized obesity and cancer information to help them make informed health decisions.

Creating KBs has been an active research area with academic projects such as YAGO [17] and NELL [18], community-driven efforts such as Freebase [19] and Wikidata [20], and commercial projects such as those by Google [21] and Facebook [22]. All of these KBs used a formal ontology-driven knowledge representation, using the Resource Description Framework (RDF) and the Web Ontology Language (OWL) to form the Semantic Web. Regardless of the debate on what is an ontology [23], an ontology can be used to encode the knowledge (i.e., specifically, the logical structure) of a domain at the semantic-level under a controlled, standardized vocabulary for describing entities and the semantic relationships between them. For example, a statement “obesity is a risk factor for cancer.” can be decomposed into a semantic triple, where the relationship between two entities“obesity” and “cancer” is “is a risk factor for”. An ontology-driven knowledge base provides a consistent representation of the facts about the domain and makes it possible to reason about those facts and use rules or other forms of logic to deduce new facts or highlight inconsistencies [24].

In the biomedical domain, there also have been several efforts in making Semantic Web KBs, such as the “Gene Ontology knowledgebase” [25, 26], UniProt [27], Literome [28], BioNELL [29], and SemMedDB [30]. These KBs provided a huge collection of entities and facts, but nonetheless, contained noises and can be unreliable due to the limited accuracy of their information extraction (IE) systems. For example, similar to our effort, SemMedDB was created with SemRep [31] based on the Unified Medical Language System (UMLS) to extract semantic triples. As of June, 2017, SemMedDB contained 91.6 million triples extracted from 27.2 million PubMed citations. Nevertheless, unlike SemRep, our approach can accurately identify meaningful entities beyond the limitation of having to be in existing terminologies. Further, recent studies have reported numerous quality issues (e.g., inaccurate facts) in the SemMedDB [32–34], for instance, “Insuli-AUGMENTS-catecholamine secretion” [35]. A major challenge in KB construction is to assess the quality of the candidate facts extracted by the IE system, and turn them into useful knowledge [36].

Crowdsourcing is a way to outsource a task to a group or community of people [37]. Crowdsourcing allows to annotate enormous volumes of data in a short time, and opened a new door to tackle complex problems and challenges in research [38, 39]. KBs can also be curated via crowdsourcing. Wikipedia (although not a Semantic Web KB) is the classic example of massively decentralized, peer-produced

KB on the Web. In terms of Semantic Web KBs, Freebase [40] and Wikidata [41] are two successful crowd-sourced examples in the general domain. The use of crowdsourcing to solve important problems in biomedical domain is also growing and has a wide variety of applications [42]. Relevant to this project, crowdsourcing is successful in biomedical natural language processing (NLP) especially on named-entity recognition [43–45], curating clinical ontology [46, 47], constructing specialized biomedical KBs [48, 49].

In our previous work [50], we created a IE software pipeline, Obesity and Cancer to Knowledge Base (OC-2-KB), aiming to build an evidence-based obesity and cancer Semantic Web KB. Using a set of NLP and machine learning techniques [51], OC-2-KB was able to automatically extract subject-predicate-object semantic triples from PubMed abstracts related to obesity and cancer. In this paper, we present a new release of the OC-2-KB system (available at: https://github.com/bianjiang/BioText-2-KB), and investigated the use of crowdsourcing to improve the quality of the extracted facts. Further, the crowd-sourced semantic triples are fed back to the machine learning modules to enrich the training corpus and subsequently improve the performance of the IE system.

The primary contributions of this work in comparison to our previous study [50], are detailed below:

- We conducted two rounds of crowdsourcing experiments to assess the feasibility of adding human knowledge to the KB construction process.
- We extended our original corpus from 23 random obesity and cancer related Pubmed abstracts to 82 abstracts from systematic review papers. We curated a preliminary obesity and cancer knowledge base (OCKB) with these abstracts using OC-2-KB and evaluated the query results against this OCKB.
- Finally, we refined the underlying machine learning models incorporating the crowdsourcing results.

The rest of the paper is organized as follows. We will briefly recap the OC-2-KB framework, and the crowdsourcing approach in the “Methods” section. The results of the crowdsourcing experiments and our evaluations of the updated machine learning models will be presented in the “Results” section. We will discuss the lessons learned, conclude the current work, and present future directions in the “Discussion and Conclusions” section.

Methods

The obesity and cancer to knowledge base (OC-2-KB) information extraction pipeline

Figure 1 shows the overview of the new release of OC-2-KB, where we added a new step called “Crowdsourcing Feedback” to the original OC-2-KB system. OC-2-KB

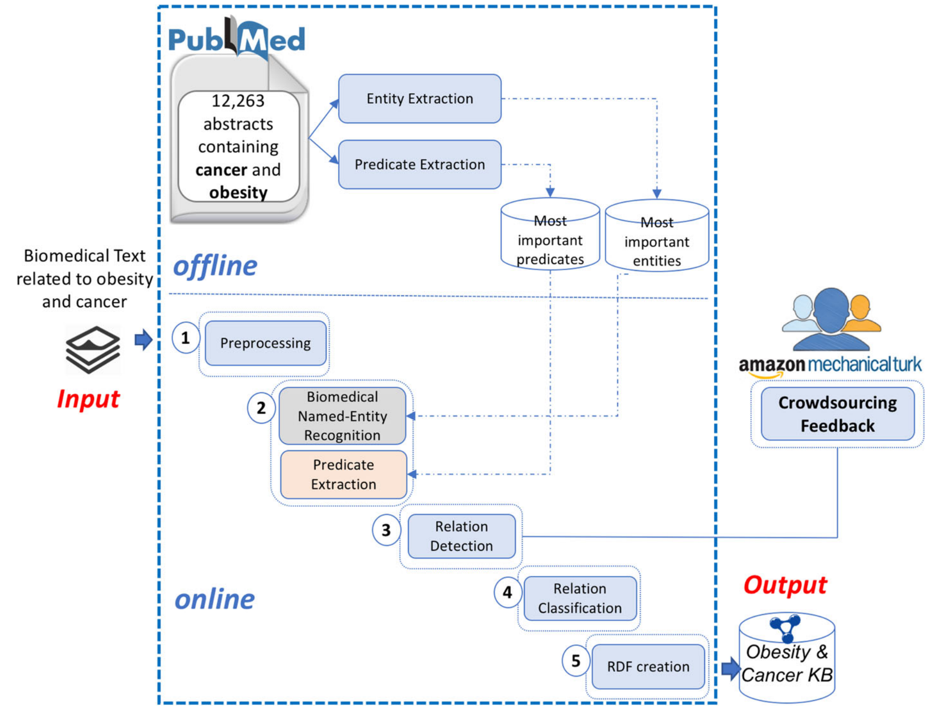


Fig. 1 An overview of the new OC-2-KB information extraction system with crowdsourcing feedback. The *offline* module creates dictionaries of domain-relevant entities and predicates. The *online* module extracts facts from scientific literature to construct an obesity and cancer knowledge base

has two components: (1) an offline process that identifies the most domain-relevant entities and predicates with importance scores, and (2) an online process that can automatically extract the facts in the format of semantic triples from relevant scientific text. For the sake of completeness, we briefly describe the crucial components of the OC-2-KB system along with our new experiments. Please refer to [50] for more system details.

The offline process: identify domain-relevant entities and predicates

The offline process was part of the configuration process of the OC-2-KB system. As our goal is to create a KB related to obesity and cancer, we used the titles and abstracts of PubMed articles containing the keywords "obesity" and "cancer" to create the dictionaries. A total of 12,263 articles were extracted. We used the *LIDF-value* [52] measure implemented in *BioTex* [53] to assess the importance of the candidate entities and predicates. Two separate dictionaries were created, which contained approximately 34,500 entities and 8,200 predicates, respectively. The entity and predicate dictionaries were then be used as inputs to the biomedical named-entity recognition (BioNER) and predicate extraction process as described below.

The online process: extract semantic triples from scientific literature

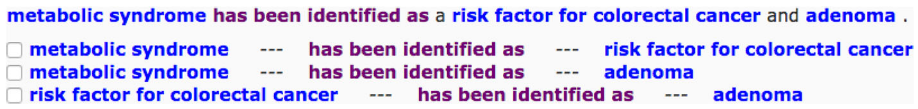
The online process has 5 main steps to extract facts in the format of semantic triples from scientific literature related to obesity and cancer.

- ***Input:*** PubMed titles and abstracts relevant to obesity and cancer.
- **Step 1 - Preprocessing:** Using the Stanford CoreNLP toolkit, we first preprocessed the PubMed titles and abstracts to split each document into a collection of sentences (i.e., sentence segmentation) as the semantic triples are currently extracted from each sentence independently.
- **Step 2 - Biomedical Named-Entity Recognition (BioNER) and Predicate Extraction:** Based on the entity and predicate dictionaries created in the offline process, we then extracted candidate entities and predicates from each sentence. As discussed in our previous work [51], our methods for biomedical NER and predicate extraction were based on both linguistic and statistic features of the text, and outperformed other state-of-the-art systems on our obesity and cancer corpus. Our experiments obtained *F-measures* of 90.1 and 51.8% for entity and predicate extractions, respectively.
- **Step 3 - Relation Detection:** Using supervised classifiers with a set of statistical, lexical, and semantic features [51], we were able to determine whether a pair of two candidate biomedical entities and a candidate predicate formed a valid assertion as a subject-predicate-object statement. The random forest algorithm achieved the best performance with an *F-measure* of 84.7% on an independent test set.
- **Step 4 - Relation Classification:** After relation detection, we normalized the predicate to one of thirteen relation classes using a multi-class classifier. Based on a manually annotated sample corpus (i.e., 23 obesity and cancer abstracts), we previously adopted [50] 12 relation classes from the Relation Ontology (RO) [54]. We added an "other" category in this work to temporarily hold all predicates that cannot be classified into one of the 12 specific relation classes. With the same set of statistical, lexical, and semantic features [51], the random forest model achieved the best results with an *F-measure* of 85.3% on a independent test set.
- **Step 5 - RDF Graph Creation:** The semantic triples extracted in the previous steps were then inserted into a graph database (i.e., GraphDB [55]) using the RDF4J [56] framework .

Incorporating crowdsourcing feedback

Even though our machine learning models demonstrated state-of-the-art performance, they were not perfect. Thus, we investigated a crowdsourcing strategy to validate the extracted semantic triples. For this purpose, we used Amazon Mechanical Turk (MTurk) [57]—an online platform that allows users to create "crowdsourcing tasks", called Human-Intelligence Tasks (HITs), to be completed for small incentives. The Amazon MTurk platform allows anyone to become a requester in order to create HITs, while people that perform those tasks are called workers [58]. Recent studies that used Amazon MTurk have shown that it is an efficient, reliable, and cost-effective research tool for generating sample responses where human feedback is essential [59, 60].

We conducted two rounds of the crowdsourcing experiment. First, a pilot study was done in September, 2017 with a small sample (i.e., one HIT that contains 10 sentences) to establish the feasibility of using crowdsourcing workers to validate the candidate triples. The small sample was randomly selected from the corpus that we used to develop the original OC-2-KB system. All candidate triples for the 10 sentences were annotated by two members of the study team (the percent agreement was 80.41%). As shown in Fig. 2, each item of the HIT was a multiple choice question, where the question represented a sentence, and each option was a candidate triple statement describing the potential relationship between two



metabolic syndrome has been identified as a risk factor for colorectal cancer and adenoma .

□ metabolic syndrome --- has been identified as --- risk factor for colorectal cancer
□ metabolic syndrome --- has been identified as --- adenoma
□ risk factor for colorectal cancer --- has been identified as --- adenoma

Fig. 2 An example question of the crowdsourcing task. The terms in blue are biomedical entities, and terms in purple are the predicates describing the relations

biomedical entities and a predicate, as "*biomedical entity 1 — relation —- biomedical entity 2*", extracted from the sentence. Workers were asked to select the most appropriate triples statements that they thought were part of the sentence.

After the pilot study, in October 2017, we launched a bigger experiment with a data set of 82 abstracts from PubMed based on the same search keywords "obesity" and "cancer". Two additional criteria were used to select the 82 abstracts: (1) articles that were review articles; and (2) articles that were published in reputable journals based on their impact factors published by Thomson Reuters. The 82 abstracts generated 671 sentences and 9,406 candidate triples. This data set is publicly available at https://github.com/bianjiang/BioText-2-KB/.

Based on experience from the pilot study, each MTurk HIT only contained 25 sentences. For each sentence, only at most the top 4 entities and top 3 predicates were extracted based on the *LIDF-value* [52] measure. We then used all possible combinations of the extracted entities and predicates as choices (i.e., candidate triples) to be validated by the workers. In sum, we created 27 HITs. Each question could be answered by 5 workers at most.

Evaluations

We performed several experiments to (1) evaluate the crowdsourcing results, (2) used the crowdsourced triples to retrain the machine learning models considering a variety of crowdsourcing parameters (e.g., the number of times each triple was validated and how long workers spent on the HITs), and (3) evaluate the KBs created with- and without- the crowdsourcing feedback.

Results

The crowdsourcing experiments

The pilot study

The pilot study was carried out in eight days, from August 31, 2017 to September 7, 2017. We created only one HIT that contained 10 sentences, and the 10 sentences were annotated by two members of the study team. The number of assignments for the HIT was set to be 400 on MTurk, i.e., 400 workers could participate at most. The reward per assignment was set to be $0.15. The allotted time for the assignment was 900 s (i.e., the worker can hold to the task for a maximum of 900 s). A total of 193 workers participated in the pilot study. Table 1 shows the configuration and cost of the MTurk HIT for the pilot study.

We also evaluated the time spent by the workers on the HIT, as shown in Table 2. We then evaluated worker performance in comparison to the gold-standard annotated by the experts, varying the amount of time spent by the workers. Table 3 shows the *F-measure* of each sentence.

Table 4 illustrates the overall worker performance in the pilot study (i.e., the average *F-measure (F)* of 10 sentences), varying by the time spent on the task. The best overall *F-measure* was 76.17%, when we accepted only answerers from workers that spent at least 300 seconds in the HIT.

The final study

For the final study, we made our HITs available on MTurk for ten days, from October 12, 2017 to October 22, 2017. Our data set was 82 abstracts extracted from PubMed. These abstracts generated 671 sentences. The maximum number of entities and predicates extracted per sentence were 21 and 11, respectively. The maximum number of possible triples was 840 (i.e., obtained from a sentence with 16 entities and 7 predicates). As crowdsourcing tasks are typically micro-tasks [61], we selected the 4 most important entities and the 3 most important predicates for each sentence according to the *LIDF-value* [52], which yielded at most 18 triples per sentence. There were a total of 9406 possible triples extracted from the 671 sentences. We then created 27 MTurk HITs, where 26 HITs contained 25 sentences and 1 HIT contained 21 sentences (a total of 671 sentences). Each HIT was set to be completed by 5 workers at most (5 assignments). The reward per assignment was $0.5. The maximal allotted time for

Table 1 Configurations of the pilot crowdsourcing study

Data	Number of assignments per HIT	400
	Reward per assignment	$0.15
Estimated	Total reward	$60.0 (= 400 × $0.15)
	Fees to Mechanical Turk	$24.0 (= 400 × $0.6)
	Total cost	$84.0
Actual Cost	Assignments done and approved	193
	Total reward	$28.95
	Fees to Mechanical Turk	$11.58
	Total cost	$40.53

Table 2 The time spent by the workers on the HIT

	Time spent on the HIT		
Time	≥ 0 s	≥ 120 s	≥ 300 s
Minimum (min)	38	127	302
Maximum (max)	889	889	889
Average ($\bar{x}$)	358.47	372.62	466.59
Standard Deviation (σ)	176.13	167.93	140.73

each assignment was 35 min (2100 s). Table 5 shows the configuration and cost of the 27 MTurk HITs. A total of 101 unique workers participated and completed 135 assignments (27 HITs × 5 assignments). As shown in Table 6, 89 workers completed only one HIT, while there is one worker completed 8 HITs.

We also evaluated the time spent by the workers, as shown in Table 7.

In addition to consider the different amount of time spent by the workers, we also considered the number of workers who made the same choices. Out of the 9406 possible triples, 3672 triples were not selected by any workers (i.e., none of the work who worked on the corresponding HIT considered the triple as valid). Table 8 presents the number of triples validated by more than 3, 4, and 5 workers (times) considering the different amount of time spent by the workers. Figure 3 shows an example of triples validated by different number of workers (i.e., more than 1, 2, 3, 4, and 5 times).

We also used the baseline OC-2-KB system (i.e., where the relation detection machine learning model was trained using the initial 23 annotated abstracts) to extract triples from the 82 abstracts *(note that the baseline OC-2-KB system extracted 765 from the 23 abstracts and 4,391 from the 82 abstracts for which only 29% of the initial facts overlapped).* We considered different confidence scores (i.e., the probability that a triple is predicted to be true by the model). Figure 4 shows the number of triples extracted by the baseline OC-2-KB system varying the confidence score (threshold λ from 0.80 to 1.00) of the relation detection machine learning model.

Table 3 Worker performance (F-measures) on the 10 sentences of the HIT

	Time spent on the HIT		
Sentences	≥ 0 s	≥ 120 s	≥ 300 s
Sentence 1	59.33%	60.24%	65.43%
Sentence 2	73.63%	75.03%	79.62%
Sentence 3	63.86%	65.30%	67.10%
Sentence 4	84.97%	84.78%	87.83%
Sentence 5	79.79%	80.62%	83.04%
Sentence 6	98.45%	98.37%	100.00%
Sentence 7	72.99%	74.02%	76.82%
Sentence 8	70.99%	72.67%	76.99%
Sentence 9	43.10%	43.30%	47.31%
Sentence 10	72.11%	73.54%	77.60%

Table 4 Overall worker performance in the pilot study

Time spent	Number of workers	F
≥ 0 s (0 s per sentence)	193	71.92%
≥ 120 s (12 s per sentence)	184	72.79%
≥ 300 s (30 s per sentence)	115	76.17 %

We compared the predicted results obtained from the baseline OC-2-KB system with the crowdsourcing results for the same 82 abstracts, as shown in Fig. 5. In this comparison, we took into account the 918 triples that were validated by at least 3 workers without the time spent constraint. Table 9 shows the number of common triples between the baseline OC-2-KB system and the crowdsourcing results ($B \cap C$), as well as the number of triples missed by the baseline OC-2-KB ($C - B$). The lowest rate of triples missed by the baseline system was 75.4% $\left(=\frac{692}{918}\right)$ with a low threshold ($\lambda = 0.80$), while the highest rate of missed triples was 92.3% with a $\lambda = 1.00$. As shown in Table 9, the baseline system missed a large number of triples, proving the necessity to augment the baseline model with human feedback.

Retraining of the relation detection model with crowdsourcing feedback

Leveraging the human feedback, we used the crowdsourced annotations to improve the machine learning models in the baseline OC-2-KB system. We considered the different combinations of the crowdsourcing parameters (i.e., considering the different number of times each triple was validated, and the different amount of time spent by the workers) to extract the positive samples as shown in Table 8, and trained 9 relation detection models with the Random Forest algorithm. The random

Table 5 Configuration and price information of the final study

Data	Number of HITs	27
	Number of assignments per HIT	5
	Reward per assignment	\$0.5
Estimated Cost	Total reward per HIT	\$2.5 (= 5 × \$0.5)
	Fees to Mechanical Turk per HIT	\$0.5 (= 5 × \$0.1)
	Total cost per HIT	\$3.0 (= \$2.5 + \$0.5)
	Total cost for 27 HITs	\$81.0 (= \$27 × \$3.0)
Actual Cost	Assignments done and approved	135 (= 27 HITS × 5 assignments)
	Total cost	\$81.0

Table 6 The number of HITs completed by the workers

Number of HITs	Number of workers that completed
1 HIT	89 workers
2 HITs	5 workers
3 HITs	1 worker
4 HITs	3 workers
5 HITs	0 worker
6 HITs	1 worker
7 HITs	1 worker
8 HITs	1 worker

forest models achieved the best performance in our previous study for the same tasks [51]. We considered the triples that were not validated by any worker (i.e., 3672) as negative samples.

Our dataset was imbalanced, since we had significant more negative samples (i.e., 3672) than positive samples (i.e., ranging from 15 to 918, as shown in Table 8). Thus, we used weighted performance metrics to evaluate the trained machine learning models. Table 10 presents the weighted F-measures of the retrained RF models for relation detection considering the different crowdsourcing parameters.

Evaluations of the knowledge bases created by the OC-2-KB system with and without the crowdsourcing feedback

To further assess the improvement of the retrained OC-2-KB system, we compared the KBs curated with the baseline OC-2-KB and retrained OC-2-KB (with the retrained model considering triples validated by more than 3 workers).

We assessed the content of the KBs (baseline OC-2-KB vs. retrained OC-2-KB) through evaluating whether the KB can answer specific competency questions. The competency questions were expressed in SPARQL queries. Note that the dataset in our original 23 abstracts was a balanced dataset (343 positive samples and 343 negative samples). To make a fair comparison, we randomly selected 918 negative samples from the 3672 triples that were not selected by any workers in the crowdsourcing experiment. In short, our training data for this part of the study were 1265 (347 + 918) positive samples and 1265 negative samples. The F-measure of the retrained relation detection model was 86.2% (ROCAUC: 0.935). For both the baseline and retrained OC-2-KB systems, we considered a threshold confidence score of 0.94.

Table 7 Time spent by the workers over the 27 HITs

	Time spent on HITs		
Time	≥ 0 s	≥ 300 s	≥ 750 s
min	43	332	761
max	2095	2095	2095
$\bar{x}$	1,001.21	1,255.35	1,421.11
σ	638.27	495.58	382.54
Number of assignments completed*	5734	4413	3770

*The number assignments completed within the time range. Note that, there were 27 HITs, and each HIT was completed by 5 workers. Thus, there are a total of 135 assignments (27 HITs times 5 workers)

Table 8 Number of triples validated more than 3, 4, and 5 times varying by the workers' time spent on the HITs

	Time spent on HITs		
The number of times validated	≥ 0 s	≥ 300 s	≥ 750 s
Validated = 5 times	37	19	15
Validated ≥ 4 times	258	109	68
Validated ≥ 3 times	918	506	320

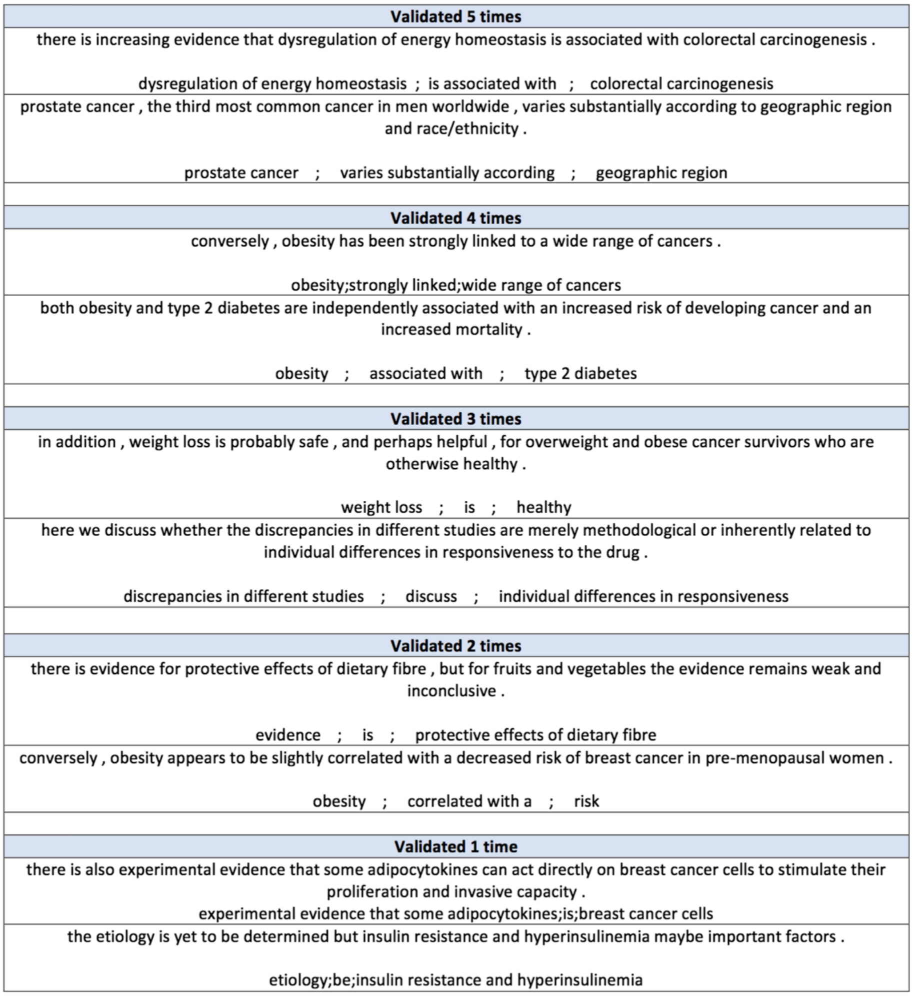
Figure 7 shows an example of all the entities related to "breast cancer risk" (i.e., entities that can form the following assertions: <oc:breast cancer risk-?predicate-?object> or <?subject-?predicate-oc:breast cancer risk>). Note that in SPARQL, a query variable is marked by the use of "?". Figure 7 (1) shows the associated query. Figure 7 (2) shows the query results from the KB created with the baseline OC-2-KB system. Figure 7 (3) illustrates the query results from the KB created with the retrained OC-2-KB system. Even though the results from the baseline contains more entities, the results from the retrained OC-2-KB are more accurate.

Figure 6 shows the receiver operating characteristic (ROC) curves of the retrained models.

Discussion

Feasibility and benefits of crowdsourcing

The non-expert crowdsourcing workers in the pilot study produced annotations comparable to the experts (i.e., an F-measure of 76.17%, which demonstrated the feasibility to use the crowd in annotation tasks. More importantly, even though our corpus was highly technical abstracts collected from scientific biomedical literature, with a careful design, cross-validated collective wisdom from laypeople can lead to quality knowledge collections. Thus, we used crowdsourcing to validate extractions that the machine identified as likely candidates. Further, validated extractions were then re-introduced into the training set for the machine learning models. As shown in Table 10 and Fig. 6, incorporating the crowdsourcing feedback significantly improved the performance of our machine learning models. Our approach is similar to the concept of active learning [62], a special case of semi-supervised machine learning where a learning algorithm is able to query the



Validated 5 times

there is increasing evidence that dysregulation of energy homeostasis is associated with colorectal carcinogenesis .

dysregulation of energy homeostasis ; is associated with ; colorectal carcinogenesis

prostate cancer , the third most common cancer in men worldwide , varies substantially according to geographic region and race/ethnicity .

prostate cancer ; varies substantially according ; geographic region

Validated 4 times

conversely , obesity has been strongly linked to a wide range of cancers .

obesity;strongly linked;wide range of cancers

both obesity and type 2 diabetes are independently associated with an increased risk of developing cancer and an increased mortality .

obesity ; associated with ; type 2 diabetes

Validated 3 times

in addition , weight loss is probably safe , and perhaps helpful , for overweight and obese cancer survivors who are otherwise healthy .

weight loss ; is ; healthy

here we discuss whether the discrepancies in different studies are merely methodological or inherently related to individual differences in responsiveness to the drug .

discrepancies in different studies ; discuss ; individual differences in responsiveness

Validated 2 times

there is evidence for protective effects of dietary fibre , but for fruits and vegetables the evidence remains weak and inconclusive .

evidence ; is ; protective effects of dietary fibre

conversely , obesity appears to be slightly correlated with a decreased risk of breast cancer in pre-menopausal women .

obesity ; correlated with a ; risk

Validated 1 time

there is also experimental evidence that some adipocytokines can act directly on breast cancer cells to stimulate their proliferation and invasive capacity .
experimental evidence that some adipocytokines;is;breast cancer cells

the etiology is yet to be determined but insulin resistance and hyperinsulinemia maybe important factors .

etiology;be;insulin resistance and hyperinsulinemia

Fig. 3 An example of triples validated through crowdsourcing

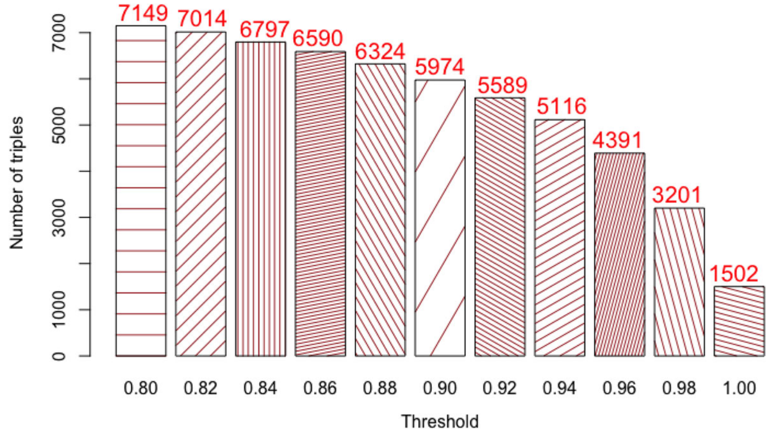


Fig. 4 The number of triples created with the baseline OC-2-KB system varying the threshold λ from 0.80 to 1.00

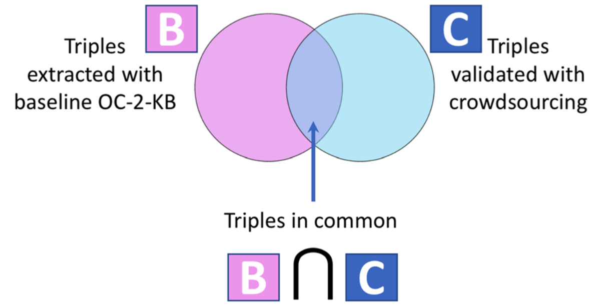


Fig. 5 A Venn diagram of comparing the number of triples extracted with the baseline system (*B*) and validated through crowdsourcing (*C*), where the number of common triples between *B* and *C* is B ∩ C

user to annotate new data points and incorporate the human feedback back into the machine learning model to improve its performance. However, as a pilot study, we selected new data to be labeled manually, while in a true active learning system, the system will be able to select unlabeled data and query the humans intelligently. This is indeed one of our future directions to create a more scalable KB construction framework.

Further, the costs of crowdsourcing studies are rather inexpensive, which is one of the other advantages of using crowd wisdom. Our pilot study received feedback from 193 workers and the cost was merely $40.53. Even if we only considered high quality responses (i.e., only accept answers from workers who spent more than 30 seconds on each sentence), 115 fell into this category. Further, the performance of these laypeople workers was comparable to expert annotators (i.e., an F-measure of 76.17%). In the final study, 89 workers completed 135 assignments (671 sentences) in 10 days, with a cost of $81. The low costs of crowdsourcing allowed us to collect redundant responses (i.e., each sentence was validated by 5 workers), which were then used to ensure the quality of the validated triples. Considering the majority rule, in our experiment, 918 triples were confirmed by at least 3 out of the 5 workers who worked on the same sentences.

Table 9 The number of common triples between the baseline OC-2-KB system and the crowdsourcing (B ∩ C), and the number of triples missed by the baseline OC-2-KB (C − B)

OC-2-KB$_\lambda$	B ∩ C	C − B
λ = 0.80	226	692
λ = 0.82	222	696
λ = 0.84	216	702
λ = 0.86	215	703
λ = 0.88	208	710
λ = 0.90	198	720
λ = 0.92	185	733
λ = 0.94	171	749
λ = 0.96	154	764
λ = 0.98	119	799
λ = 1.00	71	847

Table 10 F-measures of the retrained random forest models varying the crowdsourcing parameters

	Workers' time spent on HITs		
Number of times validated	≥ 0 s	≥ 300 s	≥ 750 s
Validated = 5 times	98.5%	98.8%	99.8%
Validated ≥ 4 times	90.3%	97.0%	97.2%
Validated ≥ 3 times	79.1%	85.1%	91.4%

The total cost of both studies on Amazon MTurk was $121.53. Amazon MTurk is proved to be an inexpensive way of gathering labeled annotations with human judgments.

The quality of the created obesity and cancer knowledge base

There is a growing body of work on automated KB construction and knowledge extraction from text. Nevertheless, the value of these KBs lies in their ability to answer users' questions (i.e., queries to the KBs). We thus evaluated the query results obtained from the OC-2-KB system with and without incorporating the crowdsourcing feedback. As shown in Fig. 7, the baseline OC-2-KB contained more but low-quality or unclear triples. For example, as shown in Fig. 7, the baseline OC-2-KB system incorrectly asserted a relation between "contrast" and "breast cancer risk". On the other hand, the triples extracted by the retrained OC-2-KB system were of higher quality. However, it is also clear that the retrained OC-2-KB system missed some of the valid triples (e.g., the relation between "family history" and "breast cancer risks"), while eliminating more false positives.

There were also triples identified by the retrained OC-2-KB system, but not by the baseline OC-2-KB system. Figure 8 shows a SPARQL query to extract any objects related to "interleukin 6". The baseline OC-2-KB system did not return any results, while the retrained OC-2-KB extracted the assertion that <"interleukin 6"-"oc:associated with"-"prostate cancer risk">.

We also assessed the predicates of triples from both KBs (baseline vs retrained). We found that in most cases the KB created with the baseline OC-2-KB extracted either incorrect predicates or semantically similar but syntactically different predicates for the same two entities. For example, Fig. 9 shows a SPARQL query to identify all relations between "progesterone levels" and "risk of endometrial cancer". The baseline OC-2-KB contained 3 predicates: "oc:associated with", "oc:is a", and "oc:other

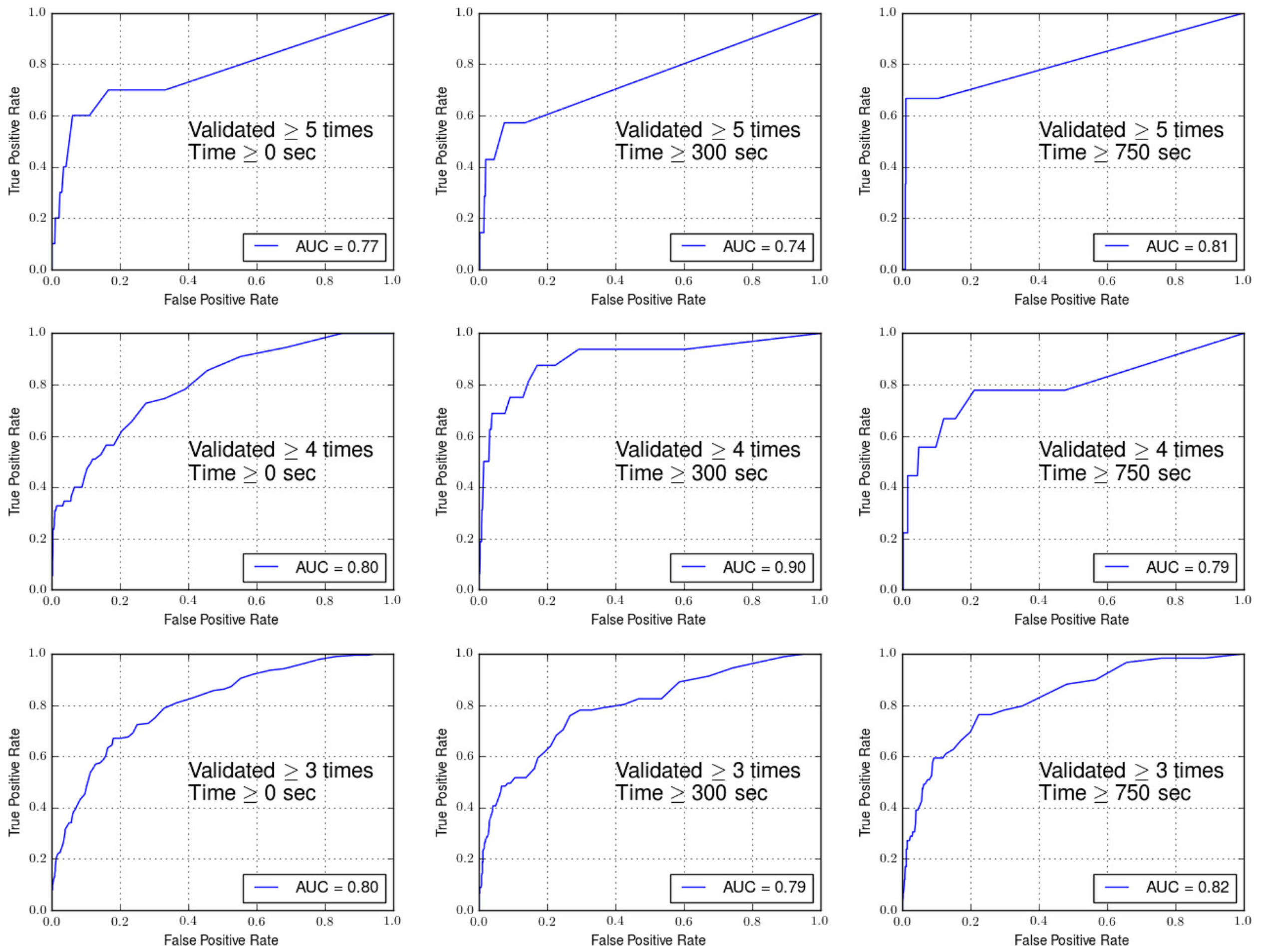
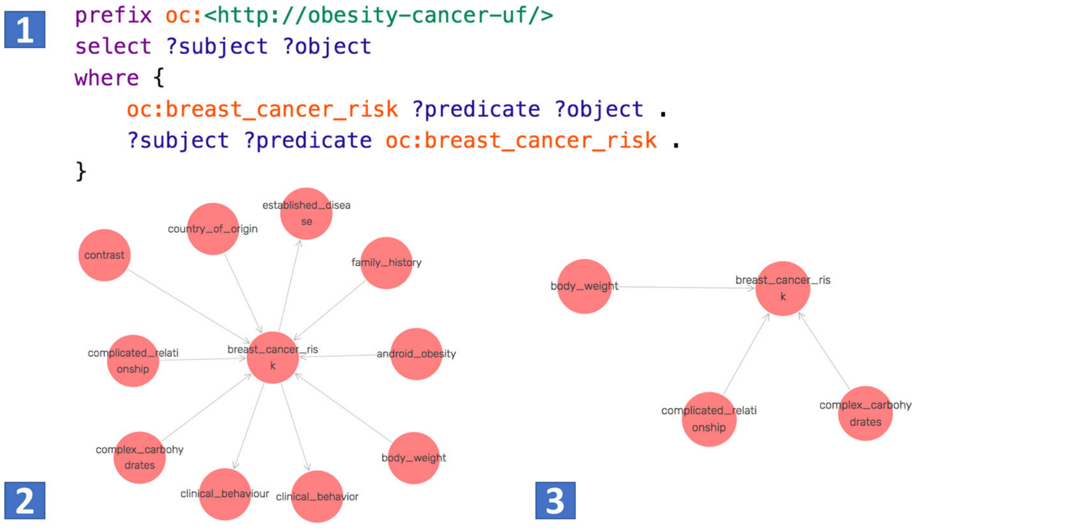
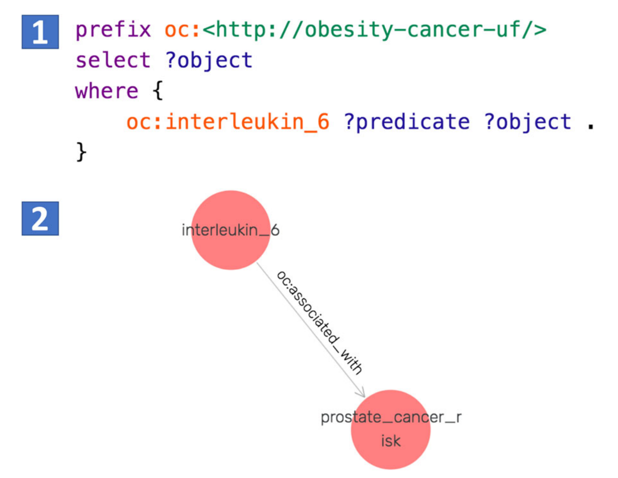


Fig. 6 The receiver operating characteristic (ROC) curves of the retrained relation detection models



```
prefix oc:<http://obesity-cancer-uf/>
select ?subject ?object
where {
    oc:breast_cancer_risk ?predicate ?object .
    ?subject ?predicate oc:breast_cancer_risk .
}
```

Fig. 7 A SPARQL query for extracting all the entities related to "breast cancer risk" in (1). The results from the baseline OC-2-KB system are shown in (2), and the results from the retrained system are in (3)



```
prefix oc:<http://obesity-cancer-uf/>
select ?object
where {
    oc:interleukin_6 ?predicate ?object .
}
```

Fig. 8 A SPARQL query for extracting entities related to “interleukin 6”

(increased)”, while the retrained OC-2-KB only contained one predicate “oc:associated with” between the same two entities.

The curated KBs, although improved with crowdsourcing feedback, still contained invalid or not meaningful triples. This is because that the accuracy of our automated relation extraction methods although achieved state-of-the-art performance is still suboptimal. Figure 10 shows an example of invalid triples related to “prostate cancer risk in men”. From the original sentence, *(“Similarly , a case-control study found obesity was inversely associated with prostate cancer risk in men aged 40-64 years.”)*, OC-2-KB extracted <“case-control study” oc:was inversely associated “prostate cancer risk in men”>, which is incorrect. This emphasizes the needs to consider humans in the loop and to use crowd wisdom to validate triples identified by the machine.

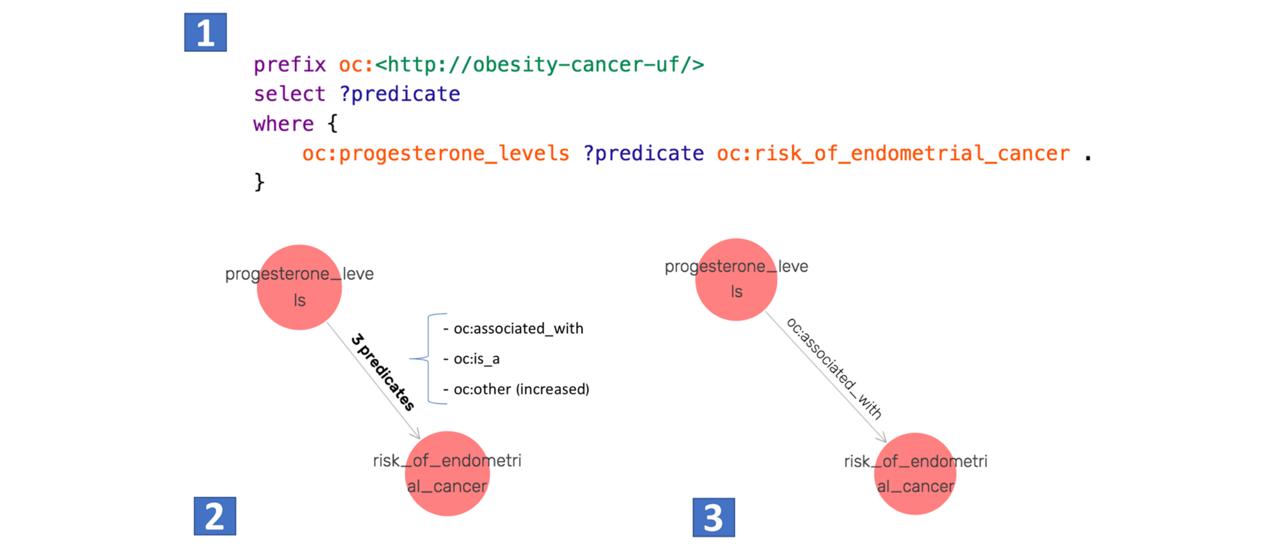
Conclusions

We presented a new version of OC-2-KB, a system that automatically builds an evidence-based obesity and cancer KB from scientific literature. We developed a scalable framework for the construction of the KB by integrating automatic information extraction with crowdsourcing techniques to verify the extracted knowledge.

Crowdsourcing platforms such as Amazon MTurk offers a low-cost mechanism to collect more labeled data from the crowd with non-expert laypeople. Even though individual layperson might not offer reliable answers, the collective wisdom of the crowd is comparable to expert opinions.

However, further studies are still warranted, as the performance of the underlying information extraction system in OC-2-KB is still suboptimal, possibly due to the noisy nature of free-text data (e.g., inconsistent formatting, alternative spellings, and misspells). We shall further investigate how to better leverage the crowdsourcing platforms. For example, we shall further explore the active learning concept, where 1) the triples that need to be validated by the crowd are identified by the machine learning model, and 2) both the precision and recall of the machine learning model will be improved with the crowdsourcing feedback, without losing existing valid triples.

The ultimate goal of this project is a paradigm shift in how the general public access, read, digest, and use online health information. Rather than requiring the laypeople find and read static documents on the Internet via regular searches, we propose a dynamic knowledge acquisition model, where the content is routinely mined



```
prefix oc:<http://obesity-cancer-uf/>
select ?predicate
where {
    oc:progesterone_levels ?predicate oc:risk_of_endometrial_cancer .
}
```

Fig. 9 Results of all predicates existing between “progesterone levels” and “risk of endometrial cancer”

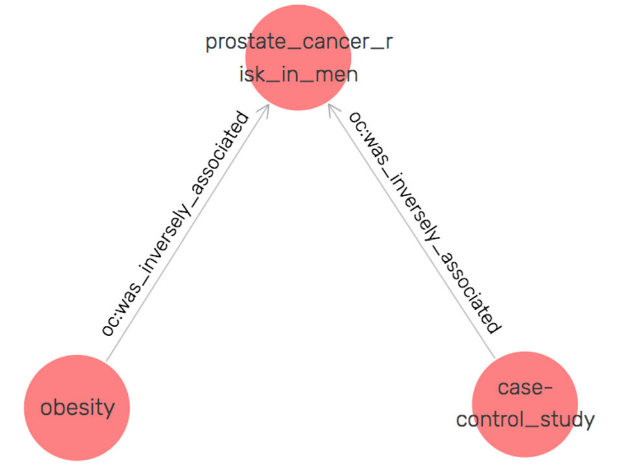

prostate_cancer_risk_in_men
oc:was_inversely_associated
oc:was_inversely_associated
obesity
case-control_study

Fig. 10 Example of invalid triples associated with "prostate cancer risk in men"

from the scientific literature and vetted, users interact via semantic queries instead of regular searches, and consumers navigate the network of knowledge through interactive visualizations.

Abbreviations

F: F-measure; IE: Information Extraction; ML: Machine Learning; NER: Named-entity Recognition; NLP: Natural Language Processing; OC-2-KB: Obesity and Cancer to Knowledge Base; OWL: Web Ontology Language; RDF: Resource Description Framework; RE: Relation Extraction; RF: Random Forest; RO: Relation Ontology; ROCAUC: Receiver operating characteristic area under curve; SPARQL: SPARQL Protocol and RDF Query Language

Funding

The work was supported in part by the OneFlorida Cancer Control Alliance (funded by James and Esther King Biomedical Research Program, Florida Department of Health Grant Number 4KB16), and the OneFlorida Clinical Research Consortium Clinical Data Research Network funded by the Patient Centered Outcomes Research Institute (PCORI). The content is solely the responsibility of the authors and does not necessarily represent the official views of UFOAPF or PCORI. Publication of this article was supported in part by the University of Florida Open Access Publishing Fund (UFOAPF).

Availability of data and materials

The annotated corpus can be found at: https://github.com/bianjiang/BioText-2-KB/tree/master/data.

About this supplement

This article has been published as part of *BMC Medical Informatics and Decision Making* Volume 18 Supplement 2, 2018: Selected extended articles from the 2nd International Workshop on Semantics-Powered Data Analytics. The full contents of the supplement are available online at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-18-supplement-2.

Authors' contributions

JALV collected the data, wrote the initial draft and revised subsequent versions. JALV and JB developed the method and performed the evaluation. JALV and JB annotated the initial corpus. JALV, JB, XY, and HZ contributed codes to the new release of the system. WH, FM, YG, and ZH contributed to the design of the study, and provided significant feedback. All authors read, revised and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, 2004 Mowry Road, 32610 Gainesville, FL, USA. [2]School of Information, Florida State University, 142 Collegiate Loop, 32306 Tallahassee, FL, USA.

Published: 23 July 2018

References

1. Organization WH, et al. Global Status Report on Noncommunicable Diseases 2014. Geneva: World Health Organization; 2014.
2. Keum N, Greenwood DC, Lee DH, Kim R, Aune D, Ju W, Hu FB, Giovannucci EL. Adult weight gain and adiposity-related cancers: a dose-response meta-analysis of prospective observational studies. J Natl Cancer Inst. 2015;107(2):088.
3. Wolin KY, Carson K, Colditz GA. Obesity and cancer. Oncologist. 2010;15(6):556–65.
4. Ligibel JA, Alfano CM, Courneya KS, Demark-Wahnefried W, Burger RA, Chlebowski RT, Fabian CJ, Gucalp A, Hershman D, Hudson MM, et al. American society of clinical oncology position statement on obesity and cancer. J Clin Oncol. 2014;32(31):3568–74.
5. Ligibel JA, Alfano CM, Hershman D, Ballard RM, Bruinooge SS, Courneya KS, Daniels EC, Demark-Wahnefried W, Frank ES, Goodwin PJ, et al. Recommendations for obesity clinical trials in cancer survivors: American society of clinical oncology statement. J Clin Oncol. 2015;33(33):3961–7.
6. Arnold M, Pandeya N, Byrnes G, Renehan AG, Stevens GA, Ezzati M, Ferlay J, Miranda JJ, Romieu I, Dikshit R, et al. Global burden of cancer attributable to high body-mass index in 2012: a population-based study. Lancet Oncol. 2015;16(1):36–46.
7. Katzke VA, Kaaks R, Kühn T. Lifestyle and cancer risk. Cancer J. 2015;21(2): 104–10.
8. Nimptsch K, Pischon T. Body fatness, related biomarkers and cancer risk: an epidemiological perspective. Horm Mol Biol Clin Investig. 2015;22(2): 39–51.
9. Fisher JD, Fisher WA. Theoretical approaches to individual-level change in hiv risk behavior. In: Handbook of HIV Prevention. Springer; 2000. p. 3–55.
10. Montano DE, Kasprzyk D. Theory of reasoned action, theory of planned behavior, and the integrated behavioral model. Health Behav Theory Res Pract. 2008;1:67–95.
11. Cantor D, Coa K, Crystal-Mansour S, Davis T, Dipko S, Sigman R. Health information national trends survey (hints) 2007 final report. 2009.
12. Purcell GP, Wilson P, Delamothe T. The quality of health information on the internet. Br Med J. 2002;324(7337):557.
13. Eysenbach G, Köhler C. How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews. Bmj. 2002;324(7337):573–7.
14. Cardel MI, Chavez S, Bian J, Peñaranda E, Miller DR, Huo T, Modave F. Accuracy of weight loss information in spanish search engine results on the internet. Obesity. 2016;24(11):2422–34.
15. Cline RJ, Haynes KM. Consumer health information seeking on the internet: the state of the art. Health Educ Res. 2001;16(6):671–92.
16. Fiksdal AS, Kumbamu A, Jadhav AS, Cocos C, Nelsen LA, Pathak J, McCormick JB. Evaluating the process of online health information searching: a qualitative approach to exploring consumer perspectives. J Med Internet Res. 2014;16(10):224.
17. Hoffart J, Suchanek FM, Berberich K, Weikum G. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. Artif Intell. 2013;194:28–61.
18. Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka Jr. ER, Mitchell TM. Toward an architecture for never-ending language learning. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence. AAAI'10, vol. 5. Palo Alto: AAAI Press; 2010. p. 1306–13. http://dl.acm.org/citation.cfm?id=2898607.2898816.
19. Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: A collaboratively created graph database for structuring human knowledge.

In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. SIGMOD '08. New York: ACM; 2008. p. 1247–50.
20. Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase. Commun ACM. 2014;57(10):78–85.
21. Dong X, Gabrilovich E, Heitz G, Horn W, Lao N, Murphy K, Strohmann T, Sun S, Zhang W. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '14. New York: ACM; 2014. p. 601–10.
22. Darwell B. Facebook builds knowledge graph with info modules on community pages. 2013. http://www.adweek.com/digital/facebook-builds-knowledge-graph-with-info-modules-on-community-pages/. Accessed 14 Jan 2013.
23. Giaretta P, Guarino N. Ontologies and knowledge bases towards a terminological clarification. Towards Very Large Knowl Bases: Knowl Build Knowl Shar. 1995;25(32):307–17.
24. Hayes-Roth F, Waterman DA, Lenat DB. Building Expert Systems. Boston: Addison-Wesley Longman Publishing Co., Inc.; 1983.
25. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25.
26. Consortium TGO. Expansion of the gene ontology knowledgebase and resources. Nucleic Acids Res. 2017;45(D1):331–8. https://doi.org/10.1093/nar/gkw1108.
27. Consortium U. Uniprot: a hub for protein information. Nucleic Acids Res. 2014;43(D1):204–12.
28. Poon H, Quirk C, DeZiel C, Heckerman D. Literome: Pubmed-scale genomic knowledge base in the cloud. Bioinformatics. 2014;30(19):2840–2.
29. Movshovitz-Attias D, Cohen W. Bootstrapping biomedical ontologies for scientific text using nell. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. BioNLP '12. Stroudsburg: Association for Computational Linguistics; 2012. p. 11–9.
30. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. Semmeddb: a pubmed-scale repository of biomedical semantic predications. Bioinformatics. 2012;28(23):3158–60.
31. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003;36(6):462–77.
32. Cameron D, Kavuluru R, Rindflesch TC, Sheth AP, Thirunarayan K, Bodenreider O. Context-driven automatic subgraph creation for literature-based discovery. J Biomed Inform. 2015;54:141–57.
33. Ayvaz S, Horn J, Hassanzadeh O, Zhu Q, Stan J, Tatonetti NP, Vilar S, Brochhausen M, Samwald M, Rastegar-Mojarad M, et al. Toward a complete dataset of drug–drug interaction information from publicly available sources. J Biomed Inform. 2015;55:206–17.
34. Shi B, Weninger T. ProjE: Embedding projection for knowledge graph completion. In: Thirty-First AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press; 2017. p. 1236–1242.
35. Zhang R, Cairelli MJ, Fiszman M, Rosemblat G, Kilicoglu H, Rindflesch TC, Pakhomov SV, Melton GB. Using semantic predications to uncover drug–drug interactions in clinical data. J Biomed Inform. 2014;49:134–47.
36. Pujara J, Miao H, Getoor L, Cohen W. Knowledge graph identification. In: Proceedings of the 12th International Semantic Web Conference - Part I. ISWC '13. New York: Springer; 2013. p. 542–57.
37. McCoy A, Wright A, Rogith D, Fathiamini S, Ottenbacher AJ, Sittig D. Development of a clinician reputation metric to identify appropriate problem-medication pairs in a crowdsourced knowledge base. J Biomed Inform. 2014;48:66–72.
38. Doan A, Ramakrishnan R, Halevy AY. Crowdsourcing systems on the world-wide web. Commun ACM. 2011;54(4):86–96. https://doi.org/10.1145/1924421.1924442.
39. Good BM, Su AI. Crowdsourcing for bioinformatics. Bioinformatics. 2013;29(16):1925–33.
40. Markoff J. Start-up aims for database to automate web searching: New York Times; 2007. http://www.nytimes.com/2007/03/09/technology/09data.html.
41. Vrandečić D, Krötzsch M. Wikidata: A free collaborative knowledgebase. Commun ACM. 2014;57(10):78–85. https://doi.org/10.1145/2629489.
42. Khare R, Good BM, Leaman R, Su AI, Lu Z. Crowdsourcing in biomedicine: challenges and opportunities. Brief Bioinform. 2015;17(1):23–32.
43. Névéol A, Doğan RI, Lu Z. Semi-automatic semantic annotation of pubmed queries: a study on quality, efficiency, satisfaction. J Biomed Inform. 2011;44(2):310–8.
44. Zhai H, Lingren T, Deleger L, Li Q, Kaiser M, Stoutenborough L, Solti I. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. J Med Internet Res. 2013;15(4):53–81.
45. Leaman R, Islamaj Doğan R, Lu Z. Dnorm: disease name normalization with pairwise learning to rank. Bioinformatics. 2013;29(22):2909–17.
46. Mortensen JM, Musen MA, Noy NF. Crowdsourcing the verification of relationships in biomedical ontologies. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association; 2013. p. 1020.
47. Mortensen JM, Minty EP, Januszyk M, Sweeney TE, Rector AL, Noy NF, Musen MA. Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of snomed ct. J Am Med Inform Assoc. 2014;22(3):640–8.
48. McCoy A, Wright A, Laxmisan A, Ottosen MJ, McCoy J, Butten D, Sittig D. Development and evaluation of a crowdsourcing methodology for knowledge base construction: identifying relationships between clinical problems and medications. J Am Med Inform Assoc. 2012;19(5):713–8.
49. McCoy A, Wright A, Krousel-Wood M, Thomas E, McCoy J, Sittig D, et al. Validation of a crowdsourcing methodology for developing a knowledge base of related problem-medication pairs. Appl Clin Inform. 2015;6(2):334–44.
50. Lossio-Ventura JA, Hogan W, Modave F, Guo Y, He Z, Hicks A, Bian J. OC-2-KB: A software pipeline to build an evidence-based obesity and cancer knowledge base. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine. BIBM'17. IEEE Computer Society; 2017. p. 1284–1287. https://doi.org/10.1109/BIBM.2017.8217845.
51. Lossio-Ventura JA, Hogan W, Modave F, Hicks A, Hanna J, Guo Y, He Z, Bian J. Towards an obesity-cancer knowledge base: Biomedical entity identification and relation detection. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Los Alamitos: IEEE; 2016. p. 1081–8. IEEE Computer Society.
52. Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M. Biomedical term extraction: overview and a new methodology. Inform Retr J. 2016;19(1-2):59–99.
53. Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M. BIOTEX: A system for biomedical terminology extraction, ranking, and validation. In: Proceedings of the 13th International Semantic Web Conference, Posters & Demonstrations Track. ISWC'14. Aachen: CEUR-WS.org; 2014. p. 157–160. http://dl.acm.org/citation.cfm?id=2878453.2878493.
54. The new OBO Relations Ontology. http://obofoundry.org/ontology/ro.html. Accessed 1 Dec 2016.
55. Ontotext GraphDB. https://ontotext.com/products/graphdb/. Accessed 20 May 2017.
56. RDF4J. http://rdf4j.org/. Accessed 10 Jan 2018.
57. Human intelligence through an API. https://www.mturk.com/. Accessed 10 January 2018.
58. Bentley FR, Daskalova N, White B. Comparing the reliability of amazon mechanical turk and survey monkey to traditional market research surveys. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. CHI EA '17. New York: ACM; 2017. p. 1092–9. https://doi.org/10.1145/3027063.3053335.
59. Chandler J, Shapiro D. Conducting clinical research using crowdsourced convenience samples. Annu Rev Clin Psychol. 2016;12:e73.
60. Mortensen K, Hughes TL. Comparing amazon's mechanical turk platform to conventional data collection methods in the health and medical research literature. J Gen Intern Med. 2018;33(4):1–6.
61. Difallah DE, Catasta M, Demartini G, Ipeirotis PG, Cudré-Mauroux P. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15. Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee; 2015. p. 238–247. https://doi.org/10.1145/2736277.2741685.
62. Settles B. Active learning literature survey. Technical report, University of Wisconsin–Madison. 2010.