


Article

Computer-Assisted Relevance Assessment: A Case Study of Updating Systematic Medical Reviews

Noha S. Tawfik ^{1,2*} and Marco Spruit ² ¹ Department of computer engineering, Arab Academy for Science and Technology, 21500 Alexandria, Egypt² Department of Information and Computing Sciences, Utrecht University, 3584 CC Utrecht, The Netherlands; M.R.Spruit@uu.nl

* Correspondence: noha.abdelsalam@aast.edu or n.s.tawfik@uu.nl

Received: 17 March 2020; Accepted: 16 April 2020; Published: 20 April 2020



Abstract: It is becoming more challenging for health professionals to keep up to date with current research. To save time, many experts perform evidence syntheses on systematic reviews instead of primary studies. Subsequently, there is a need to update reviews to include new evidence, which requires a significant amount of effort and delays the update process. These efforts can be significantly reduced by applying computer-assisted techniques to identify relevant studies. In this study, we followed a “human-in-the-loop” approach by engaging medical experts through a controlled user experiment to update systematic reviews. The primary outcome of interest was to compare the performance levels achieved when judging full abstracts versus single sentences accompanied by Natural Language Inference labels. The experiment included post-task questionnaires to collect participants’ feedback on the usability of the computer-assisted suggestions. The findings lead us to the conclusion that employing sentence-level, for relevance assessment, achieves higher recall.

Keywords: relevance assessment; information retrieval; natural language inference

1. Introduction

Relevance is a fundamental concept in Information Retrieval (IR) [1]. The entire search process revolves around the relevance concept where the effectiveness of a given system is measured by its ability to satisfy the user information needs. Defining relevance is, on its own, an active research area that started around 1959 when Brian Vickery discussed the distinction between relevance to a subject or a topic and the user relevance [2–4]. There are two main aspects of the IR process: system-driven and user-based aspects. The former focuses on finding information resources that match the user’s search query and ranking them, while the latter targets the user’s decision on assessing the relevance of the retrieved document. The main difference between both approaches is that, in the first one, the goal is to extract relevant data from a bigger pool of data, while the second’s goal is to determine the usefulness of the extracted data. This usefulness is usually subjective as it depends on the user’s information needs and varies across people [5].

A widely adopted measure of effectiveness in the IR domain is the construction of benchmark test collections on specific topics; a large document set with each document manually labeled as relevant or irrelevant. An automated process computes the similarity between IR system output and the test collection labels. This provides flexibility as system-driven evaluation can then be repeated and extended [6]. On the other hand, conducting user studies that involve human participants who test IR systems is an important method. This method has the advantage of collecting and analyzing important factors such as user judgments, user behavior, user satisfaction, and perceived problems. Despite the

accompanied issues such as the difficulty in subjects recruitment, cost and, reproducibility [7], in this study, we focused on evaluating the relevance assessment through the latter, user-based approach.

In the current era of big data, assessing relevance while maintaining a high recall is a crucial problem for information retrieval systems. In many applications, thousands of human assessments are required to achieve it, which is costly both at the time and the effort level. Examples of such applications include electronic discovery (e-discovery) especially in the legal domain, evidence-based collection in the medical domain for systematic reviews, and even construction of test collections for information retrieval tasks. High-recall driven systems are designed to reduce the cost of reliance on human input in assessing relevant documents while maintaining reasonable assessment effort.

With the exponential increase of published medical literature and the wide adoption of electronic medical records, there is an increasing need for information retrieval systems tailored to the needs of experts in searching medical textual data. Manual relevance assessment in medical IR is found to be a time demanding and cognitively expensive task [8]. Another challenge in medical IR is the variations in judgment agreements among assessors according to their expertise level, and their understanding of the document. This could profoundly impact document relevance assessment and overall retrieval performance [9]. While relevance assessment is subjective, some factors also influence the judgment accuracy, such as the search topic, the language level, the documents length, and the review time and speed.

This research aims at bridging the gap between the text mining community and the medical community. It investigates how to speed up the task of information retrieval in general, and when high levels of a recall are required specifically. We focus on one of the daily challenges in clinical practice; the relevance assessment of biomedical literature. We designed and conducted an experiment to analyze the time, effort, and decisions of medical experts in judging the relevance of scientific articles. To the best of our knowledge, there is no existing controlled user study, in the biomedical domain, that explores the relevance judgment behavior of expert assessors while varying the content-length of the previewed documents. We focused on the performance achieved by only showing the conclusion sentences paired up with computer-assisted information as opposed to showing the full abstract. The objective was to determine if acceptable recall could be achieved in a shorter amount of time by viewing less and taking advantage of extra information using language inference models. More specifically, we investigated the speed-up of the relevance assessment process, by showing less, while maintaining the quality of the judging.

2. Related Work

The research on relevance assessment involves many aspects, however, two critical factors are important for optimizing the assessment process: the time factor and the content-length factor. The time factor consists of measuring the amount of time that assessors need to perform relevance judgments, while the content-length factor denotes the percentage of text shown to users ranging from a single sentence to the full-text document.

Maddalena et al. highlighted the effect of time on the quality of the judging [10]. The authors reported that applying time constraints could reduce the costs associated with crowdsourcing with no loss of quality. Their findings show that only 25–30 s are enough to achieve top judgment quality for a single document. Wang and Soergel conducted a comparative user study in the legal e-discovery domain between participants with and without law background [11]. They investigated different parameters in relevance assessment, including speed, accuracy, and agreements. They found no significance between both groups in the speed or the accuracy when participants are provided with correct guidelines to judge document. In another study that included eight students and used the TREC e-discovery test collection, the authors conducted a correlation analysis and found that speed highly correlates to the perceived difficulty of the document [12]. However, no correlation was observed between the judgment accuracy and speed (with the exception of one topic). There was a notable variation among assessors in their judgment speed ranging from 13 to 83 documents per hour with an

average of 29 documents. It was also found that assessors judged non-relevant documents about as fast as relevant documents.

On the other hand, the effect of content-length on the relevance judgment accuracy was also investigated through user-based experiments. In [13], Tombros and Sanderson aimed at minimizing the need to refer to full-text documents by providing documents' summaries to support the retrieval decisions. Their experiment included 20 participants, and summaries were approximately 15% of the original text length mounting up to almost five sentences. The authors also introduced a 5-min time limit for each participant to identify the relevant documents. Their experimental results show that the user group that relied on summaries for judgment identified more relevant documents more accurately than the group who had access to the full document. In [14], the author investigated user-directed summaries, i.e., summaries biased towards the user's needs within the context of Information Retrieval. The paper highlights, amongst other conclusions, that users judge documents relevancy from their corresponding summaries almost as accurately as judging from their full text. It also reveals that assessing the full document needed, on average, 61 s while the document summary could be judged in only 24 s. Similarly, a study on the relevance judgment of news articles used user-biased snippets of documents in a controlled experiment. Their approach restricted shown snippets to a maximum of 50 words, an equivalent of two sentences or fewer. In that study, Smucker and Jethani found that the average time to judge a document was 15.5 and 49 s for summaries and full-documents, respectively. They also reported that the probability of judging summaries relevant is lower than documents [15]. Zhang et al. recruited 50 users to evaluate the use a single paragraph as relevance feedback in a Continuous Active Learning (CAL) settings [16]. Users were allowed only 1 h to find as many relevant documents as possible using an in-house built IR system. As expected, more relevant documents were retrieved by viewing document excerpts as opposed to full documents. In another attempt to reduce annotation overhead, the same authors suggested using single sentences for feedback in CAL [17]. The single sentence could contain sufficient information for an assessor to provide relevance judgment with an acceptable level of performance. Their results were based on a simulation study and not a human experiment, where an IR system that mimics different aspects of human interactions generates the relevance feedback.

More recently, Rahbariasl and Smucker [18] conducted a user experiment that combined different time limits with summaries and full documents. The study design included seven topics from the 2017 TREC Common Core track with 60 enrolled participants. They were given 15, 30, and 60 s of time limits to judge document relevancy with respect to a topic. The results show that, as the time limit increases, the average time to judge a document increases regardless of whether it is a full document or a summary. Overall, neither the time limits nor the document type had a statistical significant effect on accuracy. The authors suggested that employing summaries for speeding relevance judging is a better solution than imposing time limits. Their suggestion was based on the post-experiment feedback questionnaire, which shows that, with the maximum time limit of 60 s, participants enjoyed the experience with no stress involved. In that setting, they were able to judge a document summary in 13.4 s and a full-text document in 22.6 s while achieving a performance of 0.73 and 0.74 for summaries and full-text, respectively.

In the medical domain, there is scattered work on the influence of time and length variables on the assessment task based on expert judges. Existing work on relevance assessment is mostly related to TREC and CLEF evaluation challenges through the means of constructing test collections and reporting annotators' experience [19–21]. In a previous work by Koopman and Zuccon [8], the authors contradicted the findings of common intuition and previously mentioned studies as they reported that time spent to assess relevance is not related to document length but is more query-dependent. Another related study aims at investigating and improving how to effectively use search systems to extract scientific medical literature (SML) of interest. The study evaluates the impact of imposing time constraints on clinicians in answering medical questions with the help of an SML search system. The study is still ongoing, and results from participants testing is not published yet [22].

3. Materials and Methods

We assessed the effectiveness of the suggested hypothesis by means of user studies and domain test data. Generally, a test collection consists of either real or artificial data, generated by the examiner(s). We performed experimentation on a real-world case study that emulates the update process of biomedical systematic reviews (SRs) according to newly published literature.

3.1. Case Study: Updating Systematic Reviews

Scientific literature remains the primary source of information for clinicians and experts in the medical domain. The need for information synthesis is becoming indispensable, specifically with the adoption of the precision medicine concept into practice. Not only do health professionals need to keep up to date with current research, but they have to curate relevant information linking genomic variants to phenotypic data to make informed, personalized clinical decisions per patient. This makes finding relevant information more challenging, even with the availability of user-friendly search engines such as PubMed and MEDLINE that facilitate access to available publications. Additionally, the rate of publication of biomedical and health sciences literature is increasing exponentially. As an indication, it has been estimated that the number of clinical trials increased from 10 per day in 1975 to 55 in 1995 and 95 in 2015. In 2017, the PubMed repository contained around 27 million articles, 2 million medical reviews, 500,000 clinical trials, and 70,000 systematic reviews [23]. Oftentimes, medical experts do not seek answers to their questions due to time restrictions, or they suspect that no useful outcome will result from their search [24]. This tsunami of data has led time-pressured clinicians to perform evidence syntheses on systematic reviews instead of primary studies. Systematic reviews are comparative effectiveness research studies that use explicit methods to find, evaluate, and synthesize the research evidence for a research question [25]. Extracting information from reviews better suits experts' needs as there are fewer hits to curate while comparing findings across different studies efficiently [26].

Subsequently, there is a need to regularly check reviews to keep them up-to-date with current studies [26,27]. Shojania et al. [28] conducted a survival analysis on a sample of 100 systematic quantitative reviews to estimate the average time for changes in evidence. Their findings show that newly published evidence was available for 23% of the sample within two years and for 15% within one year. Moreover, 7% of the reviews were already out-of-date by the time of publication. Creating or updating systematic reviews is a time-consuming task mostly because it involves creating queries, fetching results, manual screening of candidate studies, assessing articles' relevance to the topic, and satisfying the inclusion criteria. The process of creating or updating SRs typically requires 6-12 months of effort, with the main expense being personnel time [29]. Many technology-assisted models have been proposed to reduce the efforts of the task. The models mainly rely on information extraction and text mining techniques in an automated or semi-automated approach. We designed a user-experiment to analyze the relevance assessment task within the process of updating systematic reviews. In that task, relevance assessment refers to the task of determining the relevance of an article with respect to a review topic. It can be interpreted as a measurement in which assessors judge the pertinence of a document to a chosen topic. The primary outcome of interest was the performance levels achieved when replacing the full abstract with only a single sentence enriched with textual inference labels.

3.2. Medical Natural Language Inference

This research partly builds on our previous work, in which we investigated the Natural Language Inference (NLI) task in the biomedical domain. Given two snippets of text, *Premise* and *Hypothesis*, textual inference determines if the meaning of H can be inferred from that of P [30]. Our work uses scientific literature to detect biological conflicts and literature inconsistencies between reported medical findings. The proposed NLI model follows a siamese Deep Neural Network (DNN) architecture with three layers. The input text is embedded using InferSent [31], a sentence encoder trained

on NLI sentence pairs, additional semantic features such as sentence length, cosine similarity, and contradiction-based features such as negation, modality, polarity, and antonyms. The features are integrated into the third layer of the network to improve performance. For all input pairs, the model accordingly assigns an inference label to each pair: *Entailment*, *Neutral*, *Contradictory*. The training data consist of 2135 claim pairs extracted from abstracts, on the topic of cardiovascular diseases. More details of Biomedical NLI models can be found at [32,33].

3.3. Experiment Design

The experiment session comprised two main tasks carried out in sequence, namely *Abstract-View* task and *Sentence-View* task. It also included pre and post task questionnaires for demographics information and feedback survey. The experiment was created and hosted on the online research platform Gorilla (<https://gorilla.sc/>). Gorilla is a commercial web platform originally designed for behavioral science tests through questionnaires and cognitive tests. It offers GUI construction and visual programming, as well as allows setting time limits and collection of reaction times data [34]. With these features, Gorilla was more suitable to deploy in our experiment as opposed to regular surveying tools with limited functionality when analyzing human behavior.

3.3.1. Data Collection

As mentioned, the main focus of this study was to measure how human performance changes according to the length of evidence shown. To carry such an experiment, we had to collect a set of documents with actual relevance pre-judged by domain experts. In both tasks of the experiment, users viewed a list of clinical studies associated with a systematic review extracted from the Cochrane Database of Systematic Reviews (CSDR) (<https://www.cochranelibrary.com/cdsr/about-cdsr>). The CSDR is a leading database for systematic reviews helping medical experts in the decision-making through evidence-based information. A Cochrane review identifies, appraises, and synthesizes evidence related to a specific research question given pre-specified eligibility criteria. Generally, to collect studies for systematic reviews, experts construct complex boolean queries that include logical operators such as AND, OR, and NOT to maximize the recall. The results from the query search pass through two elimination phases, abstract-screening and full-text screening. The abstract-screening involves a review of article titles and abstracts to give initial feedback on the relevancy of the study. This is followed by a full-text screening that consists of a thorough analysis to decide whether the study meets the inclusion criteria or not. Each systematic review included in the Cochrane database has its bibliography divided into three: *Included*, *Excluded*, and *Additional* sections, with references to other articles. Studies in the *Included* section are the ones found relevant to the systematic review based on the full content of the study, in other words, after passing through both elimination stages, whereas the *Excluded* section includes studies that were initially considered relevant in the abstract screening stage but later dismissed in the full-text screening stage. On the other hand, the *Additional* section includes references that are neutral and considered irrelevant to the study.

To fully assess the human judgment, the list of documents to be reviewed by the participants needs to be uniform, i.e., to maintain a good balance between relevant and irrelevant documents included in the list. Another constraint to account for while choosing the candidate documents is the difficulty level in assessing the documents; it should be as tricky as it is for experts when updating a systematic review in reality. With these goals in mind, we selected six published intervention systematic reviews, from the heart and circulation category, that are available in the CSDR. The chosen reviews are assigned the *New Search* tag, which indicates that the authors of the review published an updated version. We limited the scope of the experiment to the abstract screening stage, and hence references in the *Included* and *Excluded* sections are both considered relevant studies at the abstract level. Similarly, all references to studies in the *Additional* section are deemed irrelevant. We note that we excluded any reference to studies not related to the biomedical field, such as studies with general guidelines on how to conduct systematic reviews or interpret bias. In some cases, we re-submitted the

boolean query, if provided, to the corresponding medical database and the retrieved studies were also considered irrelevant even if not included in the *Additional* section. Our list of documents remains, however, random and un-ordered. Assessing and understanding human judgments and responses to ranked lists of documents is more complicated, and we leave it for future work. For each systematic review included in the dataset, we collected the following information: (1) review title; (2) main finding of the original systematic review; (3) PubMed ids (PMIDs) of relevant studies (references from the *included* and *excluded* sections of the updated version); and (4) PubMed ids (PMIDs) of irrelevant studies (references from the *Additional* section of the updated version or from re-submitting the query). The average percentage of relevant studies in the collected set is 45.4% of the total number of PMIDs included. Table 1 shows the distribution of the relevant and irrelevant documents of all the topics in the set.

To prepare the data for the *Abstract-View* task, PMIDs of relevant and irrelevant references were further used to download the full abstract of the studies through the Biopython library and the NCBI/PubMed API. For the *Sentence-View* task, we took advantage of the fact that abstracts of the studies follow a structured format with sections such as Background, Objectives, Results, and Conclusion. We extracted the first sentence of the Conclusion section and coupled it with the finding of the original review. The sentence pair served as input to the in-house NLI model described in Section 3.2 to generate an inference label. Additionally, each article has a relevancy value (True for relevant studies and False for not relevant) according to which reference section they belong to.

Table 1. Distribution of *Relevant* and *Irrelevant* articles in the data collection.

| Review ID | Updated Review ID | Relevant | Irrelevant | Total |
|-----------|-------------------|----------|------------|-------|
| 21249668 | 29240976 | 11 | 12 | 23 |
| 26308931 | 29667726 | 11 | 14 | 25 |
| 15846608 | 23235577 | 12 | 14 | 26 |
| 15266480 | 22293766 | 10 | 13 | 23 |
| 22696339 | 26123045 | 10 | 15 | 25 |
| 21901719 | 26934541 | 11 | 10 | 21 |

3.3.2. Participants

We aimed to recruit medical experts with knowledge in conducting systematic reviews. They were recruited by word of mouth and personal contact through emails. Aside from access to a computer or laptop with a stable Internet connection, there were no specific exclusion criteria for the study. Participants received a detailed email with the rules and information that needed to introduce the study, along with the study URL. In the invitation email, participants were asked to complete the whole experiment in a single sitting. They were also encouraged to ask for help, via email, if they had any queries or problems. All users were requested to complete a consent form prior to taking part in the experiment. Each participant completed the experiment at their convenience within two weeks from the invitation mail.

3.3.3. Interface

We plotted the experiment interface to become as much intuitive and visually pleasant as possible, and to engage participants in the document judgment task. It shows them one document at time in either tasks. For the main tasks, the screen shows either the full abstract or the conclusion sentence of the candidate study to be judged along with its corresponding PMID. We followed the standard way to collect relevance judgments where participants only could make binary judgments by clicking on relevant or irrelevant buttons. Through the whole study, participants could see the the title and original finding of the main review so that they do not forget the main topic. Figure 1 shows an example of the user interface for the main tasks.

Read the Following abstract carefully and press start to begin the experiment.

Nutritional supplementation for stable chronic obstructive pulmonary disease.

Ferreira IM, Brooks D, Lacasse Y, Goldstein RS, White J.
 PMID: 15846608

BACKGROUND: Low body weight in patients with chronic obstructive pulmonary disease (COPD) is associated with an impaired pulmonary status, reduced diaphragmatic mass, lower exercise capacity and higher mortality rate when compared to adequately nourished individuals with this disease. Nutritional support may therefore be a useful part of their comprehensive care.
OBJECTIVES: To conduct a systematic review of randomised controlled trials (RCTs) to clarify whether nutritional supplementation (caloric supplementation for at least 2 weeks) improved anthropometric measures, pulmonary function, respiratory muscle strength and functional exercise capacity in patients with stable COPD.
SEARCH STRATEGY: Randomized controlled trials (RCTs) were identified from the Cochrane Airways Group register of RCTs, a hand-search of abstracts presented at international meetings and consultation with experts. Searches are current as of March 2004.
SELECTION CRITERIA: Two reviewers independently selected trials for inclusion, assessed quality and extracted the data.
DATA COLLECTION AND ANALYSIS: Within each trial and for each outcome, we calculated an effect size. The effect sizes were then pooled by a random-effects model. Homogeneity among the effect sizes was also tested.
MAIN RESULTS: Eleven studies recruiting 352 participants met the inclusion criteria. Eight papers were considered as high quality. Two studies were double-blinded. For each of the outcomes studied, the effect of nutritional support was small: the 95% confidence intervals around the pooled effect sizes all included zero. The effect of nutritional support was homogeneous across studies.
AUTHORS' CONCLUSIONS: Nutritional support had no significant effect on anthropometric measures, lung function or exercise capacity in patients with stable COPD.

START

(a)

Original Review Title: Nutritional supplementation for stable chronic obstructive pulmonary disease

Original Review Finding: "Nutritional support had no significant effect on anthropometric measures, lung function or exercise capacity in patients with stable chronic obstructive pulmonary disease (COPD)."

PMID:19703824

Entailment This study shows that a multidisciplinary community-based disease management programme is also effective in patients with COPD with exercise impairment but less advanced airflow obstruction.

Relevant

Non-relevant

(c)

Original Review Title: Nutritional supplementation for stable chronic obstructive pulmonary disease

Original Review Finding: "Nutritional support had no significant effect on anthropometric measures, lung function or exercise capacity in patients with stable chronic obstructive pulmonary disease (COPD)."

PMID: 20150206

ABSTRACT:

Chronic obstructive pulmonary disease (COPD) is characterised by increased oxidative stress. Dietary factors, such as ample consumption of foods rich in antioxidants, such as fruit and vegetables, might have beneficial effects in COPD patients. The association between dietary shift to foods rich in antioxidants and lung function in COPD was investigated in a 3-yr prospective study. A total of 120 COPD patients were randomised to follow either a diet based on increased consumption of fresh fruit and vegetables (intervention group (IG)) or a free diet (control group (CG)). The mean consumption of foods containing antioxidants was higher in the IG than in the CG throughout the study period ($p < 0.05$). The relationship between consumption of foods rich in antioxidants and percentage predicted forced expiratory volume in 1 s was assessed using a general linear model for repeated measures; the two groups overall were different in time ($p = 0.03$), with the IG showing a better outcome. In investigating the effect of several confounders (sex, age, smoking status, comorbid conditions and exacerbation) of group response over time, nonsignificant interactions were found between confounders, group and time. These findings suggest that a dietary shift to higher-antioxidant food intake may be associated with improvement in lung function, and, in this respect, dietary interventions might be considered in COPD management.

Relevant

Non-relevant

(b)

Machine Learning suggestions

I felt the computer-assisted labels were often inaccurate

Definitely agree Somewhat agree Neutral Somewhat disagree Definitely disagree

I was confused by the computer-assisted labels

Definitely agree Somewhat agree Neutral Somewhat disagree Definitely disagree

I would like to continue using this system to aid systematic review production and update

Definitely agree Somewhat agree Neutral Somewhat disagree Definitely disagree

I found the suggested label helpful in completing the task?

Definitely agree Somewhat agree Neutral Somewhat disagree Definitely disagree

I feel that including computer-assisted information to aid reviewers would improve the quality of the final output

Definitely agree Somewhat agree Neutral Somewhat disagree Definitely disagree

Next

(d)

Figure 1. Screenshots of the user interface for both tasks and questionnaire in the experiment: (a) Original systematic review; (b) Abstract-View task; (c) Sentence-View task; and (d) Feedback questionnaire.

3.3.4. Procedures and Task Description

The designed web-based experiment for relevance assessment was divided into four main parts. Figure 2 illustrates the experiment phases.

Consent and Demographic data: At the beginning, all participants were requested to complete a consent form prior to taking part in the experiment. Next, they were asked to fill in an eleven-question survey to capture demographic data, medical background, and expertise.

Main Experiment: Following the demographic questionnaire, the main experiment started, which consisted of two tasks: the *Abstract-View* task and the *Sentence-View* task. Once the first task started, the participant reviewed specific instructions on how to complete the task. They were reminded that the experiment involves a time limit and that they should carefully read the original systematic review before moving to the assessment stage and starting the countdown. The shown instructions motivated the participants to optimize their speed and recall by instructing them to "Try to find as many relevant documents as possible in the 15 min while still making as few mistakes in judging the

documents’ relevance as possible”. Such language was also employed in similar user experiments in other domains [15,35]. First, the participant read the abstract of the original systematic review. In this step, participants acquired knowledge about the topic and also gained insights on the inclusion criteria of the review. This step was not included in the time limit since relevance judgments were based on the participants’ understanding of the topic. Next, a series of abstract texts was presented to the participant, one at a time for a duration of 15 min. To complete the task, the user selected a single judgment (Relevant, or Non-relevant). Finally, the system asked the user if this systematic review qualifies for a *Conclusion Change*, which indicates that modifications in the original findings and/or a more precise conclusion should be issued as per the Cochrane Library guidelines. The experts responded based on articles they judged as relevant and their corresponding conclusions with respect to the original SR. At the end of the first task, the system moved the participant to the second task of the experiment. The same instructions were once again shown to participants with modifications to the sentence-level task. Once again, the participant read a systematic review abstract and completed the same task in judging related documents as relevant and irrelevant. However, at that stage, the participant only viewed a single sentence and a label that indicated its inference relation with the original finding.

Each task corresponded to one of the six collected systematic reviews. We employed the built-in randomization procedure of Gorilla for assigning systematic reviews to participants. We enforced a balanced design so that all six reviews were equally included in the review and set a constraint to show different systematic review for each user per task, i.e., no duplication of documents between *Abstract-View* and *Sentence-View* tasks. At time-out, the screen was blocked, and the participant was taken to the task completion screen to enter their final details. We collected all the users’ relevance judgments and time spent to judge each document throughout both tasks. All participants acknowledged the time allocation at the start of each task but no countdown timer was visible during the task. This provided a balance between making the participant aware of the time allocated for each task without distraction that might divert their attention from from the task.

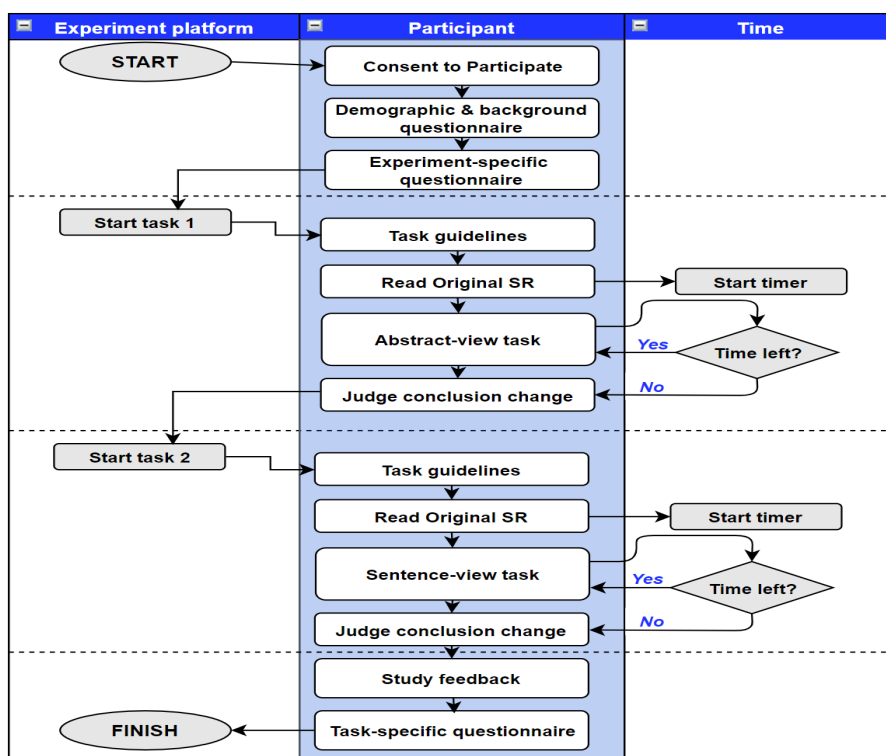


Figure 2. Process flow diagram for both tasks of the experiment.

Feedback: After completing both tasks, participants answered an exit questionnaire to collect their feedback and overall experience. To evaluate usability, we collected qualitative feedback on the experiment and the subjective perception of system usability. The feedback questionnaire follows the IBM Computer System Usability Questionnaire (CSUQ) [36]; however, the questions were adapted to fit the scope of our experiment. Additionally, we added specific questions focused on the computer-assisted suggestions to investigate the usability of the NLI independently. Finally, we asked participants to optionally provide extra feedback on the experiment via a free-text form.

Participants were able to withdraw from the study at any time by closing their browser. The whole experiment required participants to work for almost 1 h; each task required 15 min in addition to 10 min for reading the original review. We estimated that responding to the demographics and feedback survey would require 10 min. The Gorilla platform enables setting a time limit per participant to reach the final checkpoint. For convenience, we set the time limit to 90 min to allow for breaks between tasks. Participants who exceeded the time limit were eliminated from the study.

4. Results

4.1. Recruitment

Twenty-two researchers (4 male and 18 female) voluntarily participated in the study. Their age mostly fell in the 46–65 years category with 12 participants, followed by 9 participants in the 25–45 years category and only 1 participant below 25 years. All participants had masters or doctorate degrees in different biomedical fields. Most of them were affiliated to Egyptian universities, either as students (5 out of 22 participants) or as academic staff, i.e., lecturers, research assistants, or professors (17 out of 22 participants). The pre-experiment questionnaire aimed at collecting information about the participants' familiarity with biomedical literature curation and their perception of the search process. The majority of the participants use PubMed and MEDLINE (10 and 9 out of the 22 participants, respectively); alternatively, some use Google scholar for their searches (3 out of 22 participants). Table 2 summarizes the participants demographics. Almost all participants tend not to use advanced features of search engines or other computer-assisted evidence synthesis tools (95% of the participants). However, they showed interest in learning on how computer-assisted system could speed up their search. Eight participants were involved in preparing and publishing systematic reviews. As far as their perception about the different tasks for searching and validating biomedical evidence, 15 of the participants think that assessing relevance is the most challenging task, while 7 believe that the task of building queries is more complicated.

Table 2. Results of the relevance assessment experiment.

| Measure | | Sentence-View |
|------------------------------------|----------------|---------------|
| Gender | Female | 18 |
| | Male | 4 |
| Age Group | 18–24 | 1 |
| | 25–45 | 9 |
| | 46–65 | 12 |
| Education | Bachelor | 1 |
| | Master | 4 |
| | Doctoral | 17 |
| Frequency of using medical IR tool | Weekly | 15 |
| | Monthly | 6 |
| | Yearly | 1 |
| Medical IR tool | PubMed | 9 |
| | MEDLINE | 10 |
| | Google scholar | 3 |

4.2. Performance

Relevance is subjective even at the expert level, different judgments could be inferred for the same article. Accordingly, it is essential to define, within the context of the experiment, what documents count as relevant, and why. In the case study of systematic reviews update, relevance judgments are determined by the authors of the review through the reference sections, as described in Section 3.3.1. In our experiment, we considered the authors' judgments as the gold-standard annotations since they initially set the inclusion and exclusion criteria for each review. For each article, we counted documents as relevant if both the participant and the author of the systematic review agree on their relevancy, i.e., the participant clicked on the "Relevant" button, and the article is in the *included* or *excluded* reference sections, and similarly for the irrelevant documents. The correctly judged relevant and irrelevant sets are denoted S_R and S_{IRR} , respectively.

To compare the difference in relevancy judgment between sentence and abstract views, a logical measure to use is the number of correct relevant articles reported by the user, S_R , for both tasks. We interpret accuracy based on correctly assessed documents, namely the true positives (TP) and true negatives (TN), which indicate if the participants agree or disagree with the ground truth judgments. S_R represents the TP as the participant judgment aligns with the author judgment positively while S_{IRR} represents TN.

$$Accuracy = \frac{S_R + S_{IRR}}{N} \quad (1)$$

where N is the total number of articles per systematic review. Similar to the legal e-discovery task, conducting a systematic review aims at finding all available relevant documents. Missing documents could lead to legal issues for e-discovery or could affect the conclusions reported by the systematic review. Therefore, there is a need for a better suited measure of performance for this case study, such as recall, to measure the fraction of all relevant documents found by participants. In our evaluation, we calculated recall by dividing the number of correctly judged relevant documents (S_R) by the total number of relevant documents (R) available for each systematic review.

$$Recall = \frac{S_R}{R} \quad (2)$$

In statistical analysis of binary classification, the F-score, or F-measure, is a measure of a test's accuracy that combines both recall and precision to compute the score. The higher is the score, the better, reaching its best value at 1. The general formula to calculate the F-measure requires a positive real beta parameter (β) as follows:

$$F_\beta = (1 + \beta^2) \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} \quad (3)$$

Increasing the beta parameter consequently means emphasizing recall over precision. Given the recall-oriented nature of the task, we report the F2 score, which weights the recall twice as much as precision as opposed to the F1 score that portrays the harmonic mean of the precision and recall. Additionally, we tracked the time needed to judge each article in both the *Abstract-View* and *Sentence-View* tasks. As shown in Table 1, each topic has a different number of relevant documents. To minimize the skewness of reported relevant articles, we normalized each systematic review according to the number of relevant documents in its set. Table 3 shows the values for the computed measures; the corresponding values are reported on the average basis among all participants, across all systematic reviews from the data collection.

Table 3. Results of the relevance assessment experiment.

| Measure | Abstract-View | Sentence-View |
|--|---------------|---------------|
| Seconds needed to judge each article | 37.45 | 19.85 |
| Accuracy of the relevance-assessment task | 0.56 | 0.66 |
| Recall of the relevance-assessment task | 0.59 | 0.66 |
| F2-score for the relevance-assessment task | 0.57 | 0.64 |
| Number of viewed documents | 17.04 | 18.42 |
| Number of correctly judged documents | 11.00 | 12.33 |

4.2.1. Feedback

The participants found that the interface was user-friendly for the experiment and that the guidelines for each task were well explained, which made the judging tasks enjoyable. The results are based on the post-experiment questionnaire after both tasks. Almost half of the participants agreed that time limits imposed extra stress on the judgment process with an inter-participant average agreement of 2.67, while four thought that the time factor was neutral. In the second part of the survey, participants were requested to directly compare the *Abstract-View* and *Sentence-View* tasks through the following questions:

- (Q1) How difficult was it to determine if a document was relevant or not to the systematic review?
- (Q2) How would you rate your experience of judging the relevance of documents through the system?
- (Q3) How confident did you feel while doing the task?
- (Q4) How accurate do you think you have judged the documents?

We used a five-point Likert scale to map each question to values 1–5 with the most negative answer mapped to 1 and the most positive answer mapped to 5. The average response to the task-specific questions are shown in Figure 3. The results show that participants felt more confident and believe their answers were more accurate in the *Sentence-View* task. We also assessed the usability of the computer-assisted labels, based on the NLI model, in helping experts judging the relevance through six statements in the final part of the survey:

- (S1) I felt the computer-assisted labels were often inaccurate
- (S2) I was confused by the computer-assisted labels
- (S3) I would like to continue using this system to aid systematic review production and update
- (S4) I found the suggested label helpful in completing the task
- (S5) I feel that including computer-assisted information to aid reviewers would improve the quality of the final output
- (S6) I would use a similar system to to check updated information regarding a medical case

The user could express a level of agreement to the statements ranging: Definitely agree, Somewhat agree, Neutral, Somewhat disagree, and Definitely disagree. Table 4 shows the corresponding responses for each statement. The results show that participants favored the computer-assisted information and would use a similar system for curating evidence-based medicines. This is also supported by the responses to the final question in the survey where users were asked if they would recommend this system to a friend and 98% answered positively.

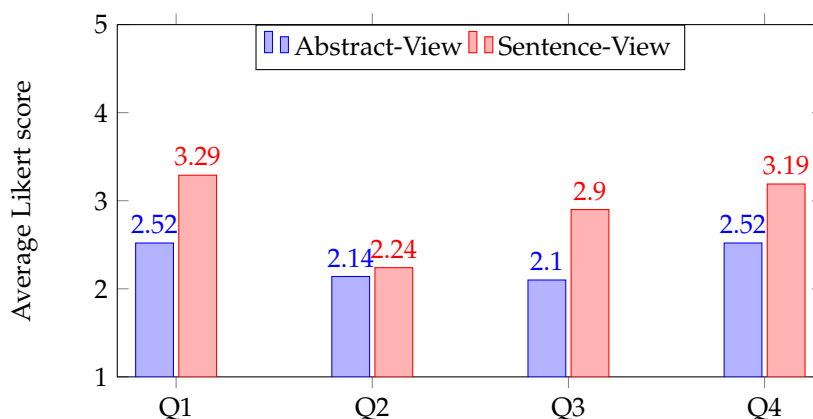


Figure 3. Task-specific questionnaire results. The participants used values 1–5 with the most negative answer mapped to 1 and the most positive answer mapped to 5.

Table 4. Participants agreement on the usability of the NLI labels.

| | Definitely Agree | Somewhat Agree | Neutral | Somewhat Disagree | Definitely Disagree |
|----|------------------|----------------|---------|-------------------|---------------------|
| S1 | 0 | 5 | 4 | 13 | 0 |
| S2 | 3 | 8 | 3 | 6 | 1 |
| S3 | 10 | 6 | 5 | 1 | 0 |
| S4 | 5 | 14 | 2 | 1 | 0 |
| S5 | 15 | 6 | 0 | 1 | 0 |
| S6 | 14 | 6 | 2 | 0 | 0 |

5. Discussion

Similar to Tombros et al. [13], we presume that the total number of experts who participated in the experiment (22 participants) is enough to draw significance to any results obtained. The following summarizes the insights from our experiment:

Less is More: There was a difference in the judging behavior among the *Abstract-View* and *Sentence-View* tasks. The assessors' quality of relevance judging when shown single sentences and labels is not only as good as when shown full abstracts but also in most the cases better in terms of accuracy and recall.

Versatile reading behavior: In the *Abstract-View* task, some participants took their time to judge the relevancy and read the abstracts carefully. Others finished the task quite soon by skimming the abstracts. This is reflected in the relatively small difference in the value of the total number of viewed articles between both tasks. We also observed that the 15 min limit provided plenty of time for participants to judge the related documents, even for the *Abstract-View* task, given the small size of the dataset (only 20–25 articles per systematic review).

The gains of the semi-automated approach: Machine learning suggestions can be used to support relevance assessment for updating systematic biomedical reviews via a web-interface. Participants appeared to engage easily with the system, rated the system as very highly usable, and would likely continue using a similar one for their daily search and curation activities. More generally, the "human-in-the-loop" experiments, such as this study, are important to demonstrate any utility afforded by text mining and machine learning.

Quality of computer-assisted suggestions: The quality of the biomedical textual inference labels is important and needs to be very accurate so that users can trust and base their judgments upon it. The feedback results show that some of the participants were confused by the highlighted inference labels. This might be because they had little knowledge on how labels were generated and were unfamiliar with the NLI model or its accuracy. A possible feature when designing an information

retrieval system is to offer an opt-in or opt-out option for showing suggestions and also providing a confidence score of the assigned labels.

6. Conclusions

In this study, we conducted a controlled user experiment to investigate the assumption that assessing document excerpts can reduce assessment time and effort and still achieve high recall. We designed a case study based on the process of updating systematic medical reviews. The case study comprises two tasks that display either full abstracts or conclusion sentences with computer-assisted labels for expert assessors to judge. Throughout each task, participants were requested to find as many relevant documents as possible within 15 min. We computed the proportion of correct judgments (i.e., the user relevancy decision is in agreement with the systematic reviews' authors) that were returned by the participants and the accuracy and recall levels in both tasks. Participants additionally completed two brief questionnaires: one at the beginning of the study, which consisted of eleven questions concerning their level of experience and demographics, and one at the end to collect the participants' feedback on the experiment in general and on the usability of the computer-assisted labels specifically. Participants were academic professionals, with a high-level of expertise, familiar with the process of conducting a systematic review. The results of the controlled user study show that assessors were able to judge more relevant articles within the time limit in the *Sentence-View* task as opposed to the *Abstract-View* task. The investigation leads us to the conclusion that employing sentence-level assessment achieves higher recall. This finding is also in alignment with previous studies [15,17,18].

Future research should extend the scope of the experiment to include participants with no medical background. This could test the hypothesis that, for the articles screening phase of the systematic review update process, a decrease in the level of knowledge of the assessor would not affect the assessment performance but would lead to a decrease in the associated costs. The main goal is to help individuals define their information needs with high precision, low burden, and minimal cost. Based on the participants' feedback, there appears to be an agreement on the usefulness of the computer-aided support in performing the task. Employing this system in other real-world cases, specifically where expert opinions are mandatory, could potentially speed up the process. The developed method could be implemented as a part of a IR system that aids medical experts in other tasks such as gathering test collections or biomedical articles curation. Another interesting future work would be adapting our system to other domains.

Author Contributions: N.S.T. and M.S. conceived and designed the experiments; N.S.T. performed the experiments, analyzed the data and drafted the paper. M.S. performed reviewing and editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Acknowledgments: We thank the volunteer researchers for participating in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mizzaro, S. Relevance: The whole history. *J. Am. Soc. Inf. Sci.* **1997**, *48*, 810–832. doi:10.1002/(SICI)1097-4571(199709)48:9<810::AID-ASI6>3.0.CO;2-U. [CrossRef]
2. Vickery, B.C. Subject analysis for information retrieval. In *Proceedings of the International Conference on Scientific Information*; The National Academies Press: Washington, DC, USA, 1959; Volume 2, pp. 855–865.
3. Vickery, B.C. The structure of information retrieval systems. In *Proceedings of the International Conference on Scientific Information*; The National Academies Press: Washington, DC, USA, 1959; Volume 2, pp. 1275–1290.
4. Saracevic, T. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *J. Am. Soc. Inf. Sci. Technol.* **2007**, *58*, 2126–2144. doi:10.1002/asi.20681. [CrossRef]

5. Janes, J.W. Other people's judgments: A comparison of users' and others' judgments of document relevance, topicality, and utility. *J. Am. Soc. Inf. Sci.* **1994**, *45*, 160–171. doi:10.1002/(SICI)1097-4571(199404)45:3<160::AID-ASI6>3.0.CO;2-4. [[CrossRef](#)]
6. Sanderson, M. Test collection based evaluation of information retrieval systems. *Found. Trends Inf. Retr.* **2010**, *4*, 247–375. doi:10.1561/1500000009. [[CrossRef](#)]
7. Kelly, D.; Kelly, D. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Found. Trends R Inf. Retr.* **2009**, *3*, 1–224. doi:10.1561/1500000012. [[CrossRef](#)]
8. Koopman, B.; Zuccon, G. Why assessing relevance in medical IR is demanding. In Proceedings of the SIGIR Workshop on Medical Information Retrieval (MEDIR 2014), Gold Coast, Australia, 11 July 2014.
9. Tamine, L.; Chouquet, C. On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. *Inf. Process. Manag.* **2017**, *53*, 332–350. doi:10.1016/j.ipm.2016.11.004. [[CrossRef](#)]
10. Maddalena, E.; Basaldella, M.; De Nart, D.; Degl'innocenti, D.; Mizzaro, S.; Demartini, G. Crowdsourcing Relevance Assessments: The Unexpected Benefits of Limiting the Time to Judge. In Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, Austin, TX, USA, 30 October–3 November 2016.
11. Wang, J.; Soergel, D. A user study of relevance judgments for e-discovery. In Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem, Pittsburgh, PA, USA, 22–27 October 2010; American Society for Information Science: Silver Springs, MD, USA, 2010; Volume 47, pp. 1–10.
12. Wang, J. Accuracy, agreement, speed, and perceived difficulty of users' relevance judgments for e-discovery. In Proceedings of SIGIR Information Retrieval for E-Discovery Workshop, Beijing, China, 28 July 2011, Volume 1.
13. Tombros, A.; Sanderson, M.; Gray, P. Advantages of query biased summaries in information retrieval. In Proceedings of the SIGIR, Melbourne Australia, 24–28 August 1998, Volume 98, pp. 2–10.
14. Sanderson, M. *Accurate User Directed Summarization from Existing Tools*; Association for Computing Machinery (ACM): New York, NY, USA, 1998, pp. 45–51. doi:10.1145/288627.288640. [[CrossRef](#)]
15. Smucker, M.D.; Jethani, C.P. Human performance and retrieval precision revisited. In Proceedings of the SIGIR 2010 Proceedings—33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva Switzerland, 19–23 July 2010; pp. 595–602. doi:10.1145/1835449.1835549. [[CrossRef](#)]
16. Zhang, H.; Abualsaud, M.; Ghelani, N.; Smucker, M.D.; Cormack, G.V.; Grossman, M.R. Effective user interaction for high-recall retrieval: Less is more. In Proceedings of the International Conference on Information and Knowledge Management, Torino, Italy, 22–26 October 2018; Association for Computing Machinery: New York, NY, USA, 2018, pp. 187–196. doi:10.1145/3269206.3271796. [[CrossRef](#)]
17. Zhang, H.; Cormack, G.V.; Grossman, M.R.; Smucker, M.D. Evaluating sentence-level relevance feedback for high-recall information retrieval. *Inf. Retr. J.* **2019**, doi:10.1007/s10791-019-09361-0. [[CrossRef](#)]
18. Rahbariasl, S.; Smucker, M.D. Time-limits and summaries for faster relevance assessing. In Proceedings of the SIGIR 2019 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; Association for Computing Machinery Inc.: New York, NY, USA, 2019, pp. 901–904. doi:10.1145/3331184.3331270. [[CrossRef](#)]
19. Kanoulas, E.; Li, D.; Azzopardi, L.; Spijker, R. CLEF 2017 technologically assisted reviews in empirical medicine overview. In Proceedings of the CEUR Workshop Proceedings, Tokyo, Japan, 11 August 2017; Volume 1866, pp. 1–29.
20. Kanoulas, E.; Li, D.; Azzopardi, L.; Spijker, R. CLEF 2019 technology assisted reviews in empirical medicine overview. In Proceedings of the CLEF 2018 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS, Avignon, France, 10–14 September 2018; pp. 1–20.
21. Roberts, K.; Demner-Fushman, D.; Voorhees, E.M.; Hersh, W.R.; Bedrick, S.; Lazar, A.J.; Pant, S. Overview of the TREC 2017 Precision Medicine Track. In Proceedings of the Twenty-Sixth Text Retrieval Conference, Gaithersburg, MD, USA, 15–17 November 2017.
22. Van Der Vegt, A.; Zuccon, G.; Koopman, B.; Deacon, A. Impact of a search engine on clinical decisions under time and system effectiveness constraints: Research protocol. *J. Med. Internet Res.* **2019**, *21*. doi:10.2196/12803. [[CrossRef](#)] [[PubMed](#)]
23. Catillon, M. *Medical Knowledge Synthesis: A Brief Overview*; NBER: Cambridge, MA, USA, 2017.

24. Ely, J.W.; Osheroff, J.A.; Chambliss, M.L.; Ebell, M.H.; Rosenbaum, M.E. Answering physicians' clinical questions: Obstacles and potential solutions. *J. Am. Med. Inf. Assoc.* **2005**, *12*, 217–224. doi:10.1197/jamia.M1608. [[CrossRef](#)] [[PubMed](#)]
25. Morton, S.; Berg, A.; Levit, L.; Eden, J. *Finding What Works in Health Care: Standards for Systematic Reviews*; National Academies Press: Washington, DC, USA, 2011.
26. Pieper, D.; Antoine, S.L.; Neugebauer, E.A.; Eikermann, M. Up-to-dateness of reviews is often neglected in overviews: A systematic review. *J. Clin. Epidemiol.* **2014**, *67*, 1302–1308. doi:10.1016/j.jclinepi.2014.08.008. [[CrossRef](#)] [[PubMed](#)]
27. Bashir, R.; Surian, D.; Dunn, A.G. Time-to-update of systematic reviews relative to the availability of new evidence. *Syst. Rev.* **2018**, *7*, 195. doi:10.1186/s13643-018-0856-9. [[CrossRef](#)] [[PubMed](#)]
28. Shojania, K.G.; Sampson, M.; Ansari, M.T.; Ji, J.; Doucette, S.; Moher, D. How quickly do systematic reviews go out of date? A survival analysis. *Ann. Intern. Med.* **2007**, *147*, 224–233. doi:10.7326/0003-4819-147-4-200708210-00179. [[CrossRef](#)] [[PubMed](#)]
29. Cohen, A.M.; Ambert, K.; McDonagh, M. A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. In Proceedings of the AMIA Annual Symposium Proceedings, Washington, DC, USA, 13–17 November 2010; American Medical Informatics Association: Bethesda, MD, USA, 2010, Volume 2010, p. 121.
30. Dagan, I.; Roth, D.; Sammons, M.; Zanzotto, F.M. Recognizing Textual Entailment: Models and Applications. *Synth. Lectures Hum. Lang. Technol.* **2013**, *6*, 1–220. doi:10.2200/S00509ED1V01Y201305HLT023. [[CrossRef](#)]
31. Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; Bordes, A. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 670–680. doi:10.18653/V1/D17-1070. [[CrossRef](#)]
32. Tawfik, N.S.; Spruit, M.R. Towards Recognition of Textual Entailment in the Biomedical Domain. In Proceedings of the International Conference on Applications of Natural Language to Information Systems, Manchester, UK, 26–28 June 2019; Springer: Manchester, UK, 2019; Volume 11608 LNCS, pp. 368–375. doi:10.1007/978-3-030-23281-8__32. [[CrossRef](#)]
33. Tawfik, N.; Spruit, M. UU_TAILS at MEDIQA 2019: Learning Textual Entailment in the Medical Domain. In Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics (ACL), Florence, Italy, 1 August 2019; pp. 493–499. doi:10.18653/v1/w19-5053. [[CrossRef](#)]
34. Anwyl-Irvine, A.L.; Massonnié, J.; Flitton, A.; Kirkham, N.; Evershed, J.K. Gorilla in our midst: An online behavioral experiment builder. *Behav. Res. Methods* **2019**, doi:10.3758/s13428-019-01237-x. [[CrossRef](#)] [[PubMed](#)]
35. Smith, C.L.; Kantor, P.B. User adaptation: Good results from poor systems. In Proceedings of the ACM SIGIR 2008—31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, 20–24 July 2008; pp. 147–154. doi:10.1145/1390334.1390362. [[CrossRef](#)]
36. Lewis, J.R. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. *Int. J. Hum. Comput. Interact.* **1995**, *7*, 57–78. doi:10.1080/10447319509526110. [[CrossRef](#)]

