Automatic Boolean Query Formulation for Systematic Review Literature Search

Harrisen Scells
University of Queensland
Brisbane, Australia
h.scells@uq.net.au

Guido Zuccon
University of Queensland
Brisbane, Australia
g.zuccon@uq.edu.au

Bevan Koopman
CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

Justin Clark
IEBH, Bond University
Gold Coast, Australia
jclark@bond.edu.au

ABSTRACT

Formulating Boolean queries for systematic review literature search is a challenging task. Commonly, queries are formulated by information specialists using the protocol specified in the review and interactions with the research team. Information specialists have in-depth experience on how to formulate queries in this domain, but may not have in-depth knowledge about the reviews' topics. Query formulation requires a significant amount of time and effort, and is performed interactively; specialists repeatedly formulate queries, attempt to validate their results, and reformulate specific Boolean clauses. In this paper, we investigate the possibility of automatically formulating a Boolean query from the systematic review protocol. We propose a novel five-step approach to automatic query formulation, specific to Boolean queries in this domain, which approximates the process by which information specialists formulate queries. In this process, we use syntax parsing to *derive the logical structure* of high-level concepts in a query, automatically *extract* and *map* concepts to entities in order to perform entity *expansion*, and finally apply *post-processing* operations (such as stemming and search filters).

Automatic query formulation for systematic review literature search has several benefits: (i) it can provide reviewers with an indication of the types of studies that will be retrieved, without the involvement of an information specialist, (ii) it can provide information specialists with an initial query to begin the formulation process, (iii) it can provide researchers that perform rapid reviews with a method to quickly perform searches.

ACM Reference Format:
Harrisen Scells, Guido Zuccon, Bevan Koopman, and Justin Clark. 2020. Automatic Boolean Query Formulation for Systematic Review Literature Search. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3366423.3380185

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380185

1 INTRODUCTION

A systematic review is a literature review that synthesises all relevant studies for a particular research question. Systematic reviews are common in the medical field, where they are important as they form the foundation of evidence based medicine, informing clinical practice, and guiding governmental and regulatory decisions.

Systematic reviews are often thought of as a *total recall* task, as they attempt to identify and synthesise all studies relevant to the review. In practice, systematic reviews attempt to achieve total recall through a variety of methods including prior knowledge of relevant studies from medical researchers, snowballing[1], and, most importantly, by formulating a *search strategy* to retrieve studies from large biomedical digital libraries (databases). Besides standard systematic reviews, other types of reviews exist where the need for total recall is relaxed, e.g., scoping reviews aim to identify key concepts and gaps within a research topic, and rapid reviews aim to provide timely information about a research question. These alternative types of reviews do not aim for total recall, but instead benefit from high precision. Nevertheless, regardless of the preference towards recall vs. precision, searching is the primary way studies are found.

A single systematic review may use multiple search strategies to find studies for inclusion. Search strategies are comprised of: (i) the database the search is submitted to, (ii) the date the search is performed (as well as any date restrictions on the publication date of studies), and most importantly, (iii) *the query* that is submitted.

Formulating a query for a systematic review, however, is a challenging task and is commonly undertaken by highly trained information specialists (e.g., specialist medical librarians). It involves constructing a complex Boolean query in order to find (all) relevant citations. In particular, the query must retrieve studies that, when synthesised in the review, contribute to the answering of the review's (generally highly focused) research question. Information specialists use domain knowledge, query formulation guidelines (e.g., [3, 9]), experience, and intuition in order to formulate queries [3].

Typically, information specialists begin the query formulation process by analysing a brief statement about the topic of the systematic review, its protocol, and a possible handful of citations

[1] Snowballing is the process by which new studies are found by examining the citations of other studies.

to clinical studies (seed citations) that the review's researchers have identified as being relevant to the review that will be conducted [3, 9]. They then identify core high-level *concepts* that emerge from the provided information. In collaboration with the researchers, information specialists then refine these concepts that form the basic structure of a query. Next, a 'proto-query' is developed and a small subset of the citations retrieved by the proto-query are screened to determine approximately how many studies may be relevant when screened. Finally, the 'proto-query' is enlarged to include synonyms of the previously identified concepts. At this stage, field restrictions (e.g., 'match only on title') may also be applied to each query keyword. The query is then 'translated' to query languages appropriate for use in a variety of medical databases to form a search strategy, e.g., into Ovid or PubMed. Different query languages cater for different advanced operators; e.g., Ovid allows for proximity operators, while PubMed does not. However, most commonly supported operators in this context are Boolean operators such as AND, OR, NOT, field restrictions (title, abstract, both), MeSH keywords[2].

Formulating queries according to the method described above can take several weeks, if not months [16, 35]. This adds to the *significant costs* (both in time and money) involved in the creation of systematic reviews. It takes in fact up to two years and a quarter of a million dollars to complete a systematic review [27]. Query formulation in this context is *demanding*, involving lengthy interactions with the search system and the underlying collection to elicit relevant terms that characterise relevant citations. Formulated queries are often *far from optimal*; i.e., they retrieve more citations than necessary (and in fact they commonly retrieve orders of magnitudes more false positives — irrelevant citations — than true positives — relevant citations), while they may still not guarantee total recall (i.e., it is unclear what the number of false negatives may be). For example, previous studies have investigated search strategies from a representative set of published high quality systematic reviews and showed that better queries than those originally in the reviews were possible; these were queries that reduced the number of false positives, while not reducing recall (or providing bounded recall losses) [38, 39]. Note that even small increases in precision can have a significant impact on both the total cost of a review, and the time required to produce a review [5, 42].

The main contribution of this paper is a computational framework for automatically formulating Boolean queries for systematic review search, which approximates the processes and intuitions of information specialists. Within our framework we present several methodologies for approximating phases of the query formulation process. Our experiments evaluate the automatically formulated queries by varying the methodologies in each step of the framework in order to identify which combination of methodologies formulates the most effective queries. We then compare these automatically formulated queries to several baselines.

2 RELATED WORK

The systematic review process comprises a number of methodological steps. Firstly, a highly specific research question is proposed which later defines the criteria by which studies should be included (and thus synthesised in the final review) and excluded from retrieved literature. Next, a search strategy is developed by one or more information specialists which attempts to capture all of the relevant studies to be included in the final systematic review. Note that these searches are usually exhaustive, leading to very small numbers of studies which meet the inclusion criteria, at times even below one percent [15]. It should also be noted that search is typically only performed on the *citations* of studies (i.e., the titles and abstracts of published full-text studies), not the full-texts. The characteristics of this type of search, which demands very high if not total recall inside a specific domain, is related to the field of 'eDiscovery' [22], which has received notable attention in Information Retrieval. In the next step, the retrieved citations are *screened* for 'potential inclusion' in the review. The full-texts of those citations that *potentially* meet the inclusion criteria are then examined to determine if they do in fact meet the inclusion criteria, or should be excluded. It is also in this phase that studies are 'snowballed' (i.e., identifying new studies by screening references of references) to find any new studies that meet the inclusion criteria, but were not retrieved by the search strategy. Finally, all of the studies which meet the inclusion criteria are synthesised and authored by the researchers into a single publication for dissemination.

Systematic reviews are costly and time consuming to create. The most expensive and time consuming aspect is the screening phase. The majority of the research that is focused on reducing the workload of the screening phase has primarily investigated the use of active learning [5, 28, 48]. Active learning techniques for document screening has received notable attention in the legal domain, where continuous active learning has been shown to significantly reduce the burden of screening on reviewers [6]. Furthermore, in this domain, neural approaches [49] to active learning have outperformed existing baseline active learning methods. These two approaches indicate the general trend outside of screening literature for systematic reviews, as well as the number of existing attempts at active learning for this domain. In addition to these works, the CLEF Technology Assisted Reviews in Empirical Methods [14] track has focused on prioritising relevant citations, and methods for determining a threshold for when to stop screening. Many submissions to this track focus on active learning. Outside work related to active learning, recent work by Lee et al. [23] has proposed a seed-driven document ranking approach, which ranks citations according to an input citation (akin to query-by-document). There has also been an uptake in the use of crowdsourcing in this setting [25, 29] as well as attention on downstream tasks, e.g., automation of results analysis [47] and synthesis automation [34, 44, 45].

Although in this paper we only focus on query formulation (i.e., search strategies), we note that the query *directly impacts how many* results are retrieved, and thus how many must be screened. Thus, savings achieved by the formulation of 'better queries' propagate throughout the systematic review creation pipeline. In this domain, two approaches to query formulation have arisen: conceptual formulation [3] and objective formulation [12]. Development of objectively derived Boolean queries compared to the conceptual approach have been found to typically yield retrieval effectiveness

[2]The **Me**dical **S**ubject **H**eadings (MeSH) ontology is an hierarchically organised index of biomedical concepts.

results of higher sensitivity [11]. However, the limitation of the objective approach is that is only applicable for meta-reviews [3] — this approach cannot be applied to typical systematic review Boolean query formulation.

The conceptual approach of query formulation begins with an information specialist identifying key high level *concepts* to construct a query given a set of citations or studies provided by the researchers conducting the review. Next, the query is enlarged with keywords which relate to the concepts. The automatic query formulation approach used in this work closely approximates the conceptual query formulation method. In addition to these two Boolean query formulation methods for systematic reviews, a number of studies have investigated 'ranked retrieval' of citations using 'textual queries' (i.e., queries similar to those found in typical ad-hoc web search tasks). Martinez et al. [26] semi-automatically created queries from the title of the review, other research questions, inclusion and exclusion criteria, and by flattening the original Boolean query. Karimi et al. [15] also created queries in a similar manner: a combination of title, background information (e.g., research questions), and inclusion criteria; as well as flattening the original Boolean query. While these studies do indeed show that using ranked retrieval without Boolean queries can reduce the workload associated with screening (although it does not deliver the recall levels expected for this task), Boolean queries are exclusively used for search in this domain. Kim et al. [17] has developed a decision tree based method for Boolean query formulation in eDiscovery, focused on query suggestions. This method uses pseudo-relevant studies in order to select which concepts to add to a Boolean query as well as the location of the concept in the Boolean query structure (i.e., by considering a binary tree as an equivalent representation of Boolean functions). This method is challenging to apply to systematic review literature search due to the difficulty in obtaining high quality pseudo-relevant studies.

The generation of structured queries from natural language information needs has been studied also in natural language processing and database research, where the resulting queries are expressed in SQL, e.g. [1, 30, 32, 50]. These methods however do not cater for the specific nature of systematic reviews, including the host of operators used in this field, the complexity of the queries, the sheer quantity of synonyms, etc..

3 THE BOOLEAN QUERY FORMULATION FRAMEWORK

In this paper, we propose a novel framework to *automatically formulate Boolean queries* for systematic review literature search. The input to the framework is a short statement about the topic of the review, e.g., "galactomannan detection for invasive aspergillosis in immunocompromised patients". This is generally created by researchers following the **P**opulation, **I**nterventions, **C**ontrols, and **O**utcomes (PICO) technique to frame and answer clinical questions, and is provided as part of the protocol of the review. In addition, the framework may be given seed citations as input; i.e., studies that the researches know a priori to be relevant. Commonly, researchers provide expert information specialists with a handful of seed citations. Our framework does not expressively require seed citations to be given, and in its current implementation it does not use seed citations. The framework could, for example, be expanded to consider seed citations for relevance feedback mechanisms within the entity expansions step.

The framework comprises five steps (Figure 1): query logic composition (1), entity extraction (2), entity expansion (3), keyword mapping (4), and post-processing (5). These steps approximate the process an information specialist undertakes when formulating queries. In *query logic composition* (step 1), a logical hierarchy of high-level concepts in a query is extracted from a short description about the systematic review (the brief topic statement). In *entity extraction* (step 2), the high-level concepts in the query are extracted and represented with entities from a reference entity repository (i.e., a medical terminology or thesaurus, e.g., UMLS[4]). In *entity expansion* (step 3), a query is broadened by adding related entities, within relevant locations in the query's logical structure. In *keyword mapping* (step 4), entities in the query are mapped to one or more keyword expressions[5]: keywords replace entities in the query's logical structure. In *post-processing* (step 5), stemming and study filters may be applied. Study filters are standard Boolean expressions developed by the systematic reviews community for explicitly retrieving specific types of citations, e.g., randomized control trials (RCTs). The third and fifth steps are optional, and valid queries can be obtained by applying step 4 immediately after step 2: this may result in narrower queries being formulated. Each of these five steps is described in further detail in the following sections.

3.1 Query Logic Composition

Once provided with a high level overview of the topic of a review, the information specialist usually proceeds by deriving the key *high-level concepts* of the review, which in turn inform the logical structure of the final query. An example of a high-level concept from the example query in Figure 1 is "Immunocompromised Patients". Ultimately, high-level concepts are used to identify which *keywords* to use in the query. In this step, the information specialist also decides the main logical structure of the query in terms of Boolean operators. Typically, the specialist groups the highest-level concepts with `AND` operators, as all these concepts need to be in relevant citations and related lower-level concepts with `OR` operators (as these keywords form alternative expressions for referring to the high-level concept)[3].

Our framework attempts to automatically approximate the extraction of the key high-level concepts from a systematic review topic statement by analysing it using an unlexicalised, English probabilistic context-free grammar (PFCG) parser [18] to segment words from systematic review statements into noun phrases. The noun phrase structure (i.e., parse tree) is then mapped directly to a Boolean query. Concepts are grouped by the presence of noun phrases: we assume that nested noun phrases indicate the semantic grouping of phrases. Noun phrases at the highest level are grouped by the Boolean operator `AND`, and noun phrases at lower depths in the query structure are grouped by the Boolean operator `OR`.

[3] A meta-review can be considered a systematic review of systematic reviews [4].

[4] The **U**nified **M**edical **L**anguage **S**ystem (UMLS) is an integration of a number of key medical and biomedical terminologies, including MeSH.

[5] Each keyword may be formed by an n-gram or phrase.

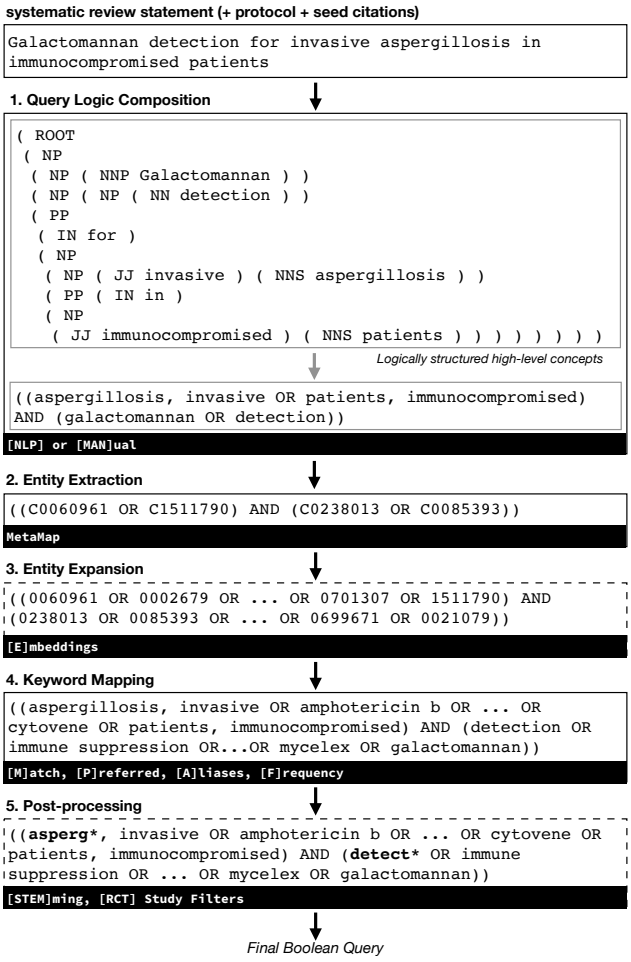

Figure 1: Our automatic Boolean query formulation framework. Three high-level concepts are identified from the systematic review's topic statement: *galactomannan detection*, *invasive aspergillosis*, and *immunocompromised patients*. These are then logically composed into a Boolean syntax and entities extracted (defined as entries in the UMLS). Entities are then (optionally) expanded. Entities are then mapped to keywords: in the example, we use the *Alias* method. Finally, (optional) post-processing is applied to obtain the Boolean query (in the example the query was stemmed).

Figure 1 provides the parse tree for the query in our example, and its mapping to a proto-query with the relevant Boolean syntax.

In our experiments (Section 4), we refer to the above method for extraction of high-level concepts and generation of the Boolean syntax as **NLP**. We shall also consider the **[Man]ual** creation of the parse tree, to study the impact of possible errors from the PFCG parser on the effectiveness of the formulated queries. The structure of the query and the selection of Boolean operators is derived from the manually created parse tree, in the same way as the NLP method.

3.2 Entity Extraction

Following the logical composition of the review's topic statement into a proto-query, entities are extracted from the query's high-level concepts. To represent entities we use the Unified Medical Language System (UMLS). [6] We use UMLS, rather than MeSH (which is commonly used as part of systematic reviews queries) as UMLS comprises a larger scope of topics than MeSH, and provides many alternative ways to refer to an entity. Note that MeSH is also subset of the UMLS. The mapping between high-level concepts and UMLS entities (also known as CUIs) is performed through any of the available medical entity extraction algorithms, e.g., MetaMap [2] or QuickUMLS [43], which use natural language processing and term matching between the keywords to express the high-level entity and the terms used to describe entities in UMLS. In our experiments we use MetaMap, and leave the studying the impact of alternative entity extraction techniques to future work.

3.3 Entity Expansion

Once the entities that form the query have been extracted, an optional expansion step can be applied to broaden the query with related entities. This step is optional — if not applied, the resulting query may, however, be narrow and thus not suited for recall-oriented searches; but may be sufficient for precision-oriented searches. To identify expansions, entities (UMLS CUIs) are embedded into a high-dimensional vector space (using word embeddings techniques [46]). The cosine similarity between each entity in the proto-query and any entity in the embedding space is computed. For each proto-query entity, the k entities with the highest cosine similarity are selected and included in the query, respecting the query's logical structure, and the entities contained in the proto-query. We classify this method of entity expansion **[E]mbedding**. Note that other methods are possible to perform entity expansion: we could consider methods of medical entity similarity surveyed by relevant literature [7, 19, 31] — we leave this for future work.

3.4 Keywords Mapping

Queries are then processed so as to transform entities (CUIs) into keywords — these could be single words or phrases. Several methods to map entities to keywords are explored in this research:

- **[M]atch**: The original keywords matched by the NLP tool to extract entities from high-level concepts are used to map back these entities to keywords and are added to the query. When using MetaMap as an NLP tool, this corresponds to using `Matched` terms. Note that this method can only be used when directly applied to the proto-query, i.e., when no entity extraction has been performed.
- **[P]referred**: The UMLS `Preferred` terms associated to an entity (CUI) are added to the query.
- **[A]liases**: An entity in UMLS is associated to one or more terms which describe aliases for that entity. All the identified terms are then added to the query.
- **[F]requent**: Rather than adding all aliases of an entity, only the most frequently used term is added to the query. The intuition for this method is derived from work by Jimmy et al. [13].

[6] Although other suitable terminologies may be used (e.g., MeSH, SNOMED CT, etc.).

3.5 Post-Processing

The result of keyword mapping is a Boolean query which can be directly executed on medical databases to search for literature. This query can, however, be further processed by applying post-processing operations. We consider two post-processing techniques: stemming, and the addition of randomised controlled trial (RCT) filters. Other activities could be performed, e.g., the addition of MeSH terms, limiting certain keywords to specific fields (title, abstract, etc.), adjusting proximity operators between keywords (depending on the query language and database used). We leave the development and study of other post-processing activities to future work. We describe the two post-processing techniques considered in this work below.

[STEM]ming: Often keywords in queries for systematic reviews are explicitly, manually stemmed to increase the possibility of matching relevant citations. The use of standard English stemmers, e.g., Porter and Krovetz [20, 33], may not be suitable to keywords in these queries as they are medical terms which are frequently derived from Latin and Greek — and may be a combination of several words [21], e.g., "acrocephalopolysyndactylie". While alternative stemming algorithms exist that may better cater for medical language [21, 41], their effectiveness in this context is still poorly documented. We then resort in taking a statistical approach to stemming. We obtain a collection of Boolean queries published in systematic reviews[7] and we extracted their stemmed keywords; in published reviews, these are indicated with wildcards, (e.g., ?, *, etc.). Then, for each keyword in an automatically formulated Boolean query, we substitute the longest match found in the list of stems, if any.

[RCT] Filters: An emerging trend among information specialists undertaking systematic review querying is the development of methodological search filters [8], usually in an attempt to increase the precision of the searches (though filters exists that aim to increase recall). Search filters are a quasi-standard Boolean expression comprising keywords and indexing terms (e.g., MeSH) that are designed to retrieve a specific type of literature. For example, filters exist for the type of studies the review seeks, e.g. Figures 2 and 3 show the sensitivity-maximising and sensitivity-and-precision-maximising search filters for randomised control trials (RCTs) studies, as developed by Cochrane initiative [9]. These are the filters we consider in our experiments, although other filters may have been applied. Specifically, in our experiment we evaluate our framework using a collection containing Diagnostic Test Accuracy (DTA) reviews: however, the use of routine filters for these reviews is discouraged [10], and standard, comprehensive filters for DTA are not available. We study the application of RCT filters as an example of the use of search filters in our framework because often DTA reviews rely on the analysis of randomised control trial studies.

[7]Including the queries for the collection we used for the experiments.

```
#1 randomized controlled trial [pt]
#2 controlled clinical trial [pt]
#3 randomized [tiab]
#4 placebo [tiab]
#5 drug therapy [sh]
#6 randomly [tiab]
#7 trial [tiab]
#8 groups [tiab]
#9 #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8
#10 animals [mh] NOT humans [mh]
#11 #9 NOT #10
```

Figure 2: Cochrane sensitivity-maximising RCTs filter.

```
#1 randomized controlled trial [pt]
#2 controlled clinical trial [pt]
#3 randomized [tiab]
#4 placebo [tiab]
#5 clinical trials as topic [mesh: noexp]
#6 randomly [tiab]
#7 trial [ti]
#8 #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7
#9 animals [mh] NOT humans [mh]
#10 #8 NOT #9
```

Figure 3: Cochrane sensitivity-and-precision-maximising RCTs filter.

4 EXPERIMENTAL SETUP

4.1 Dataset

To validate the effectiveness of our automatic Boolean query formulation framework we used the CLEF Technology Assisted Reviews (TAR) 2018 collection [14]. This collection contains queries, protocols, relevance assessments, and links to the original reviews for 72 (training and testing combined) Diagnosis Test Accuracy (DTA) reviews. For evaluation, we used the abstract-level relevance judgements (assessments from TAR-2018 Subtask 2). A relevant citation is a retrieved study that has met the inclusion criteria during the screening phase, but may have been later excluded during the appraisal of studies based on full-text. This level of relevance was chosen for the experiment because the alternative (relevant when meeting the inclusion criteria *and* included in the review) assumes knowledge about the full-text of each study (which is not used in this work). We performed the experiments on all 72 queries (none of the methods studied here required training data), by directly executing our queries on PubMed using the Entrez API [36]. The experimental pipelines for automatically formulating and evaluating queries was performed through Querylab [37], which allows for the release of formulation and evaluation pipelines in a common format for reproducibility[8].

The collection contains Boolean queries in two query languages: Ovid Medline and PubMed. As our experiments use PubMed as

[8]These will be made available in an online appendix upon publication.

the basis for retrieval, it is necessary to translate the Ovid Medline queries to the PubMed query language. Although we used an automatic tool for this translation [37], a number of queries could not be automatically translated due to formatting/logical errors, and we applied manual intervention to help in this process.

4.2 Implementation of the Framework

As input to our framework we used the title reported in the protocol associated to each systematic review in the CLEF TAR-2018 collection; this is an approximation of the systematic review brief topic statement. Note that the title of a systematic review does not contain information that would be known *after* the review is complete: from this perspective it is appropriate to use as if the researchers of a review were to provide it to an information specialist as a summary/statement about the topic of the review. Figure 1 shows an example title from the CLEF 2018 collection.

When implementing the proposed framework, we used the Stanford unlexicalised PFCG English parser[9] [18] for the **[NLP]** method in the *query logic composition* step. We also performed a manual segmentation and parsing of the title (**[Man]**), to control for errors in the automatic parsing. We did this in a consistent way: for example, when concepts relating to 'diagnosis' appeared, we always formed a new clause; the PICO method was also used to drive the segmentation of titles into high-level concepts.

To perform the *entity extraction* step, we used *MetaMap* version 2018, with options set to their default values.

The *entity expansion* step (**[E]**) relied on the availability of embeddings to find related entities (UMLS CUIs) to those in the protoquery. To this aim we used a resource containing 500k clinical concept embeddings by van der Vegt et al. [46], which learned embeddings for a large set of UMLS CUIs on all of PubMed. When considering the top similar entities to a target entity, we set $k = 20$.

Methods **[P]**, **[A]** and **[F]** relied on data about entities in the UMLS meta-thesaurus; for this, we used the UMLS 2018-AB meta-thesaurus. Method **[M]** instead relied on the n-grams from the systematic reviews' statements as extracted by MetaMap, for which we used the same Metamap instance used for entity extraction.

To derive a set of common medical stems for the **[STEM]** method, we used the Boolean queries from the CLEF 2018 collection and those in the collection by Scells et al. [40]. The latter consists of Boolean queries from a set of 125 high quality systematic reviews developed through the Cochrane initiative (and not focused on DTA, as is the case for CLEF 2018 instead).

4.3 Evaluation Measures

When evaluating the effectiveness of the automatically formulated queries, we considered two contexts: the formulation of a query for a standard systematic review, and that for a rapid systematic review. Both types are used for DTA reviews. The difference is that standard systematic reviews require high level of recall (if not total recall); while rapid reviews trade off losses in recall for high(er) precision. To model these two evaluation contexts, we measure recall, F_1, F_3 and work saved over sampling (WSS) [5] when considering standard systematic reviews; we measure precision, $F_{0.5}$ and F_1 when considering rapid reviews. We report these evaluations contexts separately, so as to not confuse the need for total recall with the trade off of recall in favour of precision.

Statistical significance analysis of the methods' comparative effectiveness is performed using two-tailed paired Student's t-test with Bonferroni correction.

[9] https://nlp.stanford.edu/software/lex-parser.shtml

	Rapid Reviews			Systematic Reviews		
	Precision	$F_{0.5}$	F_1	F_3	WSS	Recall
O	0.0253	**0.0310**	**0.0469**	**0.1578**	**0.9039**	**0.9105**
S	**0.0264**	0.0309	0.0434	0.1216	0.4927	0.4929

Table 1: Evaluation results of the original (O) queries versus the simplified (S) queries.

4.4 Baselines

To facilitate a fair comparison between our methods and the queries that reviewers manually formulated for the same reviews, the original queries formulated by human information specialist in the collection were modified for a separate baseline. The queries were modified by removing MeSH index terms and normalising all field restrictions were to `Title/Abstract`. This was done as the queries formulated using our framework do not make use of MeSH index terms or attempt to apply different field restrictions. We refer to these modified queries as *"simplified"*, this is denoted as `S`. For completeness, in Table 1 we report the evaluation results for the original queries (without modifications) and their simplified counterpart. Additionally, we consider a *naïve* approach to query formulation where Boolean queries were formed using terms from the reviews' titles connected by a Boolean OR operator; this is denoted as `T`.

We also performed experiments using the Boolean query generation algorithm proposed by Kim et al. [17] to use as a baseline, which was devised for creating queries for patent search. To adapt their method to the task of query generation for literature search, we submitted the systematic review topic description (title) to PubMed and use its ad-hoc query ranking mechanism to obtain pseudo-relevant documents. We followed this approach because their method requires as input a set of (pseudo-) relevant documents; indeed, the method is thought as a mean for further refining an initial query. The queries obtained when applying this method, however, were sub-par and did not contribute to a meaningful comparison, so were omitted from the results reported in the following section. This highlights the difficulties of adapting methods that at first appear applicable to this task; it also further demonstrates the novelty of the method proposed here.

5 RESULTS

The results of the first four steps of the automatic query formulation are presented in Table 2. Overall, there is no clear winner in terms of an automatic query formulation method that outperforms the simplified original query on all measures for these particular queries. The method with the highest F_1, however, is MAN/M: a combination of Manual query logic composition and Matched entity mapping. This method does indeed improve over the simplified baseline in terms of all rapid review oriented evaluation methods; however, it suffers significantly in terms of typical systematic review evaluation measures. Meanwhile, the MAN/F/E is not significantly worse than the simplified query in any measure,

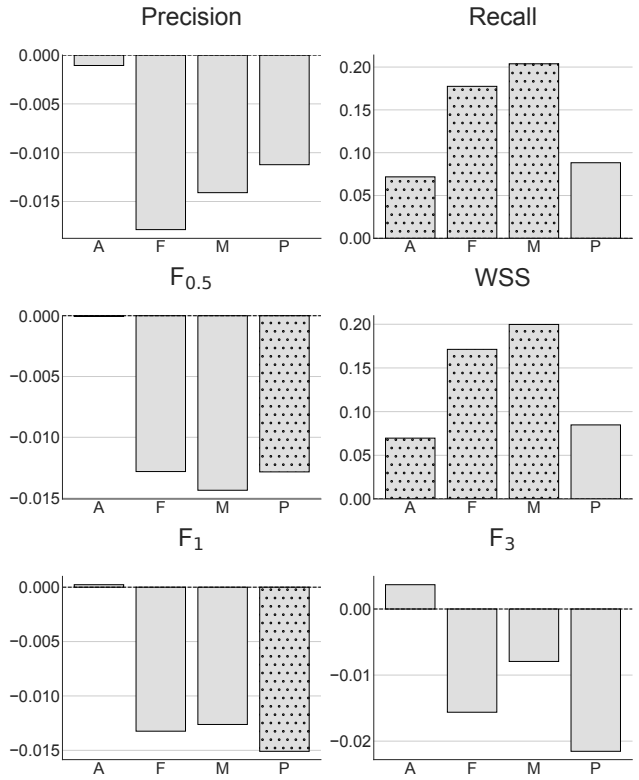

Figure 4: NLP query logic composition method versus manual method. The left-hand figures indicate measures relating to rapid reviews, while the right-hand figures indicate measures relating to typical systematic reviews. Each subfigure illustrates the gains or losses the NLP method has over the manual method. A bar in the positive y-axis indicates a gain for the NLP method over the Manual method, and a bar in the negative y-axis indicates a gain for the Manual method over the NLP method. Two-tailed statistical significance with $p < 0.01$ is indicated by bars with a ∴ pattern.

indicating that these queries better approximate how the simplified query was formulated. Note that the post-processing methods are not shown in this table: these results are discussed later. Even still, there is much variability between how queries are logically composed, which initial keywords are chosen for these queries, and even which keywords are chosen as expansion terms. The following sections explores differences between the query logic composition Manual and NLP methods; next, the impact of entity to keyword mapping, then the impact of entity expansion, and finally, the impact of post-processing.

5.1 Impact of Query Logic Composition

When considering the query logic composition methods, there exists a trade-off between the NLP and Manual methods in terms of precision and recall. Figure 4 illustrates these differences in query logic composition. Overall, the NLP method composes queries that are, on average, significantly more effective in terms of typical systematic review oriented measures (e.g., recall) over the Manual method, but less effective in terms of rapid review oriented measures (e.g. precision). Interestingly, however, are queries where concepts are eventually mapped to keywords in terms of F_3, which weights recall higher than precision: here, the Manual method composes queries that are, on average, more effective.

These results suggest that, for rapid review oriented measures, the structure of the query (i.e., how keywords are combined by Boolean operators) is more important than the keywords used. Whereas for systematic review oriented measures, the selection of keywords in the query play a larger role than the structure of the query.

5.2 Impact of Keyword Mapping

The keyword entity mapping method also produces measurably different queries between each query logic composition method and between each entity mapping method. Table 3 provides a comparison of entity mapping methods between query logic composition methods. Statistically significant differences between the entity mapping methods can be seen for both query logic composition methods. Notably, both query logic composition methods show many significant differences in terms of systematic review oriented evaluation measures for the Alias and Preferred methods. Overall, the systematic review oriented evaluation measures are most affected by the entity mapping methods. The Alias entity mapping method provides the highest recall, WSS and F_3 for the Manual query logic composition method, while the Match method provides the highest values for these evaluation measures for the NLP method.

In terms of rapid review oriented evaluation measures, there are fewer significant differences between entity mapping methods. The one exception is the Preferred mapping method for the Manual query logic composition method. This method is statistically significantly worse than all other methods. There is no statistical significant differences between the NLP queries, however the Preferred method also produces queries which are much worse than the other methods.

5.3 Impact of Entity Expansion

Each entity expansion method impacts the effectiveness of a query in different ways as well – this is visualised in Figure 5. Each subfigure shows the average difference between the evaluation scores of a query formulated query logic composition and entity mapping, and when these queries have an additional entity expansion step applied. (Note that entity expansion cannot be applied to the Match method, see Section 3). All entity expansion methods except NLP/P/E provide gains in systematic review oriented measures. Some gains in recall are statistically significant (two-tailed t-test with $p < 0.1$), however the losses in precision and F_1 are also statistically significant in some cases. Most notable is the MAN/A/E query formulation method, where an increase in effectiveness is obtained for all measures, and a statistically significant increase in effectiveness for both systematic review and rapid review oriented evaluation measures can be observed.

		Rapid Reviews			Systematic Reviews		
		Precision	$F_{0.5}$	F_1	F_3	WSS	Recall
	S	0.0264	0.0309	0.0434	0.1216	0.4929	0.4927
	T	0.0001	0.0001	0.0002	0.0011	0.9303	0.8425
Manual	M	**0.0446**$^{\triangle}$	**0.0345**$^{\triangle}$	**0.0335**$^{\triangle}$	0.0479$^{\blacktriangledown\triangle}$	0.1527$^{\blacktriangledown\triangledown}$	0.1525$^{\blacktriangledown\triangledown}$
Manual	A	0.0057$^{\blacktriangledown}$	0.0043$^{\blacktriangledown}$	0.0051$^{\blacktriangledown}$	0.0090$^{\blacktriangledown}$	0.0328$^{\blacktriangledown\triangledown}$	0.0326$^{\blacktriangledown\triangledown}$
Manual	A/E	0.0110$^{\blacktriangledown\triangle}$	0.0127$^{\blacktriangledown\triangle}$	0.0177$^{\blacktriangledown\triangle}$	**0.0553**$^{\blacktriangledown\triangle}$	0.5874$^{\triangledown}$	**0.5842**$^{\triangledown}$
Manual	F	0.0441$^{\triangle}$	0.0315$^{\triangle}$	0.0315$^{\triangle}$	0.0482$^{\blacktriangledown\triangle}$	0.1608$^{\blacktriangledown\triangledown}$	0.1604$^{\blacktriangledown\triangledown}$
Manual	F/E	0.0223$^{\triangle}$	0.0236$^{\triangle}$	0.0307$^{\triangle}$	0.0673$^{\triangle}$	0.3477$^{\triangledown}$	0.3471$^{\triangledown}$
Manual	P	0.0285$^{\triangle}$	0.0236$^{\triangle}$	0.0249$^{\triangle}$	0.0403$^{\blacktriangledown\triangle}$	0.1163$^{\blacktriangledown\triangledown}$	0.1162$^{\blacktriangledown\triangledown}$
Manual	P/E	0.0116$^{\triangle}$	0.0115$^{\triangle}$	0.0145$^{\blacktriangledown\triangle}$	0.0379$^{\blacktriangledown\triangle}$	0.2961$^{\blacktriangledown\triangledown}$	0.2947$^{\blacktriangledown\triangledown}$
NLP	M	0.0305	0.0201$^{\triangle}$	0.0208$^{\triangle}$	0.0400$^{\blacktriangledown\triangle}$	0.3565$^{\triangledown}$	0.3523$^{\triangledown}$
NLP	A	0.0046$^{\blacktriangledown}$	0.0042$^{\blacktriangledown}$	0.0053$^{\blacktriangledown}$	0.0127$^{\blacktriangledown}$	0.1045$^{\blacktriangledown\triangledown}$	0.1023$^{\blacktriangledown\triangledown}$
NLP	A/E	0.0162	0.0118$^{\blacktriangledown}$	0.0120$^{\blacktriangledown\triangle}$	0.0270$^{\blacktriangledown\triangle}$	**0.5941**$^{\triangledown}$	0.5594$^{\triangledown}$
NLP	F	0.0262$^{\triangle}$	0.0187$^{\triangle}$	0.0183$^{\triangle}$	0.0326$^{\blacktriangledown\triangle}$	0.3383$^{\triangledown}$	0.3317$^{\triangledown}$
NLP	F/E	0.0018$^{\blacktriangledown}$	0.0008$^{\blacktriangledown}$	0.0008$^{\blacktriangledown}$	0.0020$^{\blacktriangledown}$	0.0452$^{\blacktriangledown\triangledown}$	0.0451$^{\blacktriangledown\triangledown}$
NLP	P	0.0173	0.0107$^{\blacktriangledown\triangle}$	0.0099$^{\blacktriangledown\triangle}$	0.0188$^{\blacktriangledown\triangle}$	0.2046$^{\blacktriangledown\triangledown}$	0.2010$^{\blacktriangledown\triangledown}$
NLP	P/E	0.0052$^{\blacktriangledown\triangle}$	0.0055$^{\blacktriangledown\triangle}$	0.0073$^{\blacktriangledown\triangle}$	0.0186$^{\blacktriangledown\triangle}$	0.2896$^{\blacktriangledown\triangledown}$	0.2809$^{\blacktriangledown\triangledown}$

Table 2: Results of the automatic query formulation methods. Two-tailed statistical significance ($p < 0.01$) is computed between each formulation method and the original simplified (S) and title methods (T) (indicated by ▲, △ if a method is statistically significantly higher, or ▼, ▽ if a method is statistically significantly lower, respectively). Solid triangles indicate significant differences between S, outlined triangles indicate significant differences between T. Values in bold indicate the highest value out of all automatic methods.

		Rapid Reviews			Systematic Reviews		
		Precision	$F_{0.5}$	F_1	F_3	WSS	Recall
Manual	M	0.0481^{p}	0.0391	0.0388	0.0557^{a}	0.1765^{a}	0.1767^{a}
Manual	P	0.0273^{m}	0.0243	0.0261^{a}	0.0418^{a}	0.1424^{a}	0.1440^{a}
Manual	A	0.0478	0.0390	0.0459^{p}	0.0889^{mpf}	0.3565^{mpf}	0.3572^{mpf}
Manual	F	0.0429	0.0365	0.0382	0.0564^{a}	0.1672^{a}	0.1674^{a}
NLP	M	0.0305	0.0201	0.0208	0.0400	0.3523^{p}	0.3565^{pa}
NLP	P	0.0173	0.0107	0.0099	0.0188	0.2010^{maf}	0.2046^{maf}
NLP	A	0.0198	0.0154	0.0175	0.0412	0.4741^{pf}	0.4962^{mpf}
NLP	F	0.0262	0.0187	0.0183	0.0326	0.3317^{pa}	0.3383^{pa}

Table 3: Comparison between the entity mapping methods between each query logical composition method. Two-tailed statistical significance ($p < 0.01$) is computed between each entity mapping for a query logical composition method (indicated by ^{M} for the Match method, ^{P} for the Preferred method, ^{F} for the frequency method, and ^{A} for the alias method).

5.4 Impact of Post-Processing

When considering the impact of post-processing on queries, the results of the two most effective queries are chosen for analysis (NLP/A/E, and MAN/M). Figure 6 visualises the impact of RCT filters and stemming has on these chosen queries. Each sub-figure shows the average difference between the evaluation scores of the simplified original query and the automatically formulated query with either the sensitivity-maximising filter, sensitivity-and-precision-maximising filter, or stemming applied. When compared to the simplified original query, most post-processing filters saw a reduction in effectiveness (leading to why only the two best methods are chosen for analysis). These results highlight the importance of using search filters that are not only specific to the kind of systematic review (i.e., DTA reviews or rapid reviews), but to the topic of the systematic review (i.e., systematic reviews about cancer should have a cancer-specific search filter). In using unspecific or general search filters, the effectiveness of a query can be significantly degraded.

While the use of search filters is pervasive in systematic review search, our experiments show that they actually had a detrimental effect on retrieval effectiveness. This may not be the case for all

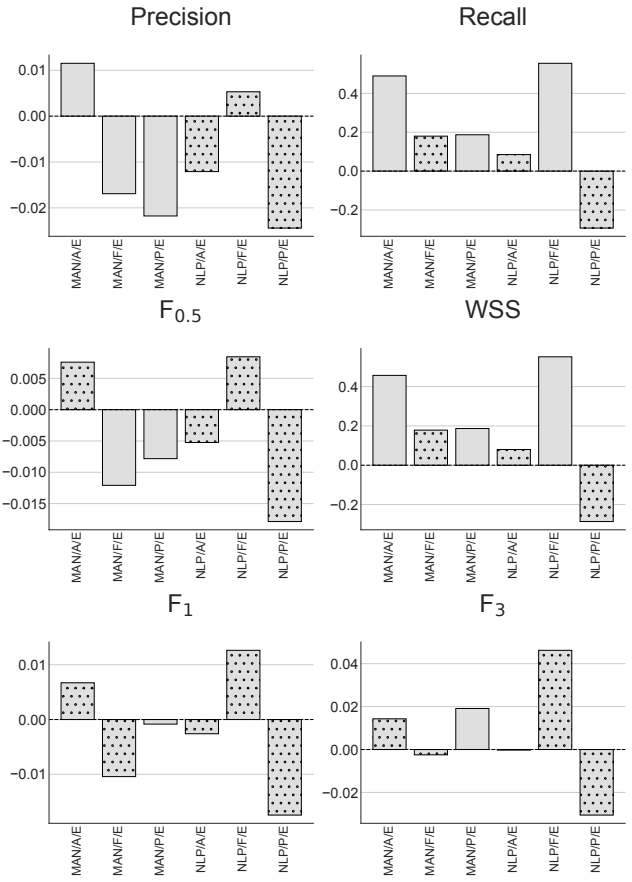

Figure 5: Impact of the entity expansion methods when applied to queries formulated using each query logic composition and entity mapping method. The left-hand figures indicate measures relating to rapid reviews, while the right-hand figures indicate measures relating to typical systematic reviews. Each sub-figure shows the average difference between the evaluation scores of the original automatically formulated query and the same query without the entity expansion applied. Two-tailed statistical significance with $p < 0.01$ is indicated by bars with a ∴ pattern.

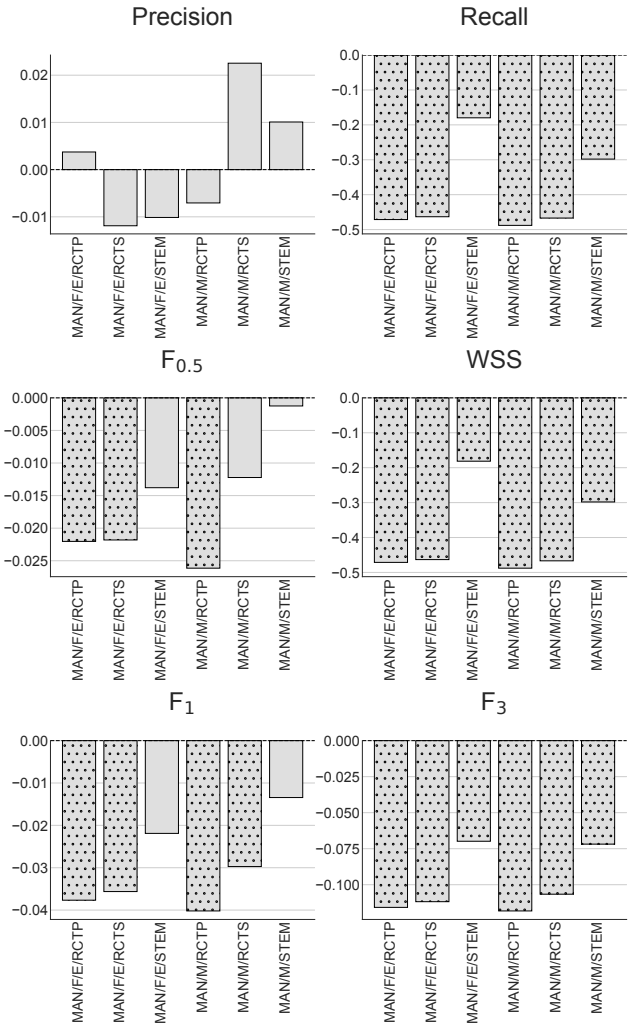

Figure 6: Impact of the post-processing on automatically formulated queries. The left-hand figures indicate measures relating to rapid reviews, while the right-hand figures indicate measures relating to typical systematic reviews. Each sub-figure shows the average difference between the evaluation scores of the human gold standard query and the automatically formulated query with either the sensitivity-maximising filter (RCTS), or the sensitivity-and-precision-maximising filter (RCTP), or stemming (STEM). Two-tailed statistical significance with $p < 0.01$ is indicated by bars with a ∴ pattern.

systematic reviews; however, the results at least serve as cautionary tale for commonly used, yet experimentally unproven, search practises.

6 DISCUSSION

The results of the automatic Boolean query formulation experiments show just how difficult the task of query formulation is for this domain. Our novel five-step approach to query formulation, however, can outperform a comparable query formulated manually (i.e., the simplified original query). There often exists a trade-off between precision and recall using these approaches, and no single combination of automatic query formulation methods was able to, on average across all considered evaluation measures, outperform the simplified original query. In terms of rapid review oriented measures, the MAN/E method automatically formulates queries with the highest gains. And in terms of systematic review oriented measures, the NLP/A/E method automatically formulates queries which are closest in performance to the baseline. These results are important as it shows that the methods laid out in this work are able to automatically formulate queries that are comparable to an equivalent query that could be formulated manually. Moreover, these methods can produce a variety of queries, automatically, which

either maximise or minimise particular evaluation measures (and thus incur a trade-off between precision and recall) or which closely approximate what could have been formulated manually — meaning that these methods can be applied to a variety of contexts.

The MAN/F/E approach to automatic query formulation, produces queries which do not suffer significant differences with the simplified original queries. This method is a promising candidate to focus on when refining the query formulation process. For example, a number of parameters about this method may be tweaked (e.g., the number of expansions made) which may further improve the effectiveness.

When adding a randomised controlled trial filter to the query, it can be observed that for these particular queries which are searching literature for diagnostic test accuracy systematic reviews, the filters significantly reduce the effectiveness of queries. This is due to the difficulty of systematic reviews on diagnostic test accuracy. Evidence based on the experience of biomedical researchers searching for diagnostic test accuracy studies for a systematic review, has shown that typical randomised controlled trial filters are not effective in this context [24]. Clearly, these general purpose RCT filters are not suitable for queries retrieving literature for systematic reviews of diagnostic test accuracy. However, the sub-par effectiveness of these queries highlight the importance of choosing appropriate search filters.

7 CONCLUSION & FUTURE WORK

There is clear room for future work in this area. Two important aspects about queries in this context were not considered in the experiments: field restrictions and MeSH keywords & explosion. We expect future work which focuses on these aspects to see considerable gains in effectiveness, and automatic formulation of queries close to the effectiveness of Boolean queries formulated manually by professionals. Extensions and additions to the query logic composition, entity mapping, and entity expansion methods as proposed in this work are also available areas for future work. For example, there are a number of possible avenues for future work and extensions that are mentioned in this paper which may help to further improve these results. However, we predict only marginal gains in effectiveness could come about.

Next, the identification of appropriate filters is necessary for queries retrieving literature for systematic review on diagnostic test accuracy. Future work on developing these filters is necessary, as it is still unclear if global filters (e.g., the quasi-standard Cochrane randomised controlled trial filters) or local filters (i.e., dependent on the query) are more effective. Finally, work can be done to identify the effectiveness of these automatic query formulation methods on broader types of systematic reviews (e.g, typical systematic reviews which synthesise randomised controlled trials, scoping reviews which aim to synthesise literature on a very broad topic, or meta-reviews which are systematic reviews on systematic reviews).

Our novel five-step approach to automatic query formulation was informed by the practices of real information specialists formulating Boolean queries for systematic review literature search. Our approach can automatically produce queries that are often as effective as an equivalent manually formulate query. Moreover, these queries require minimal data (one sentence) in order to be formulated: in our experiments only the title of the systematic review is used. Our motivation is not to replace the information specialist performing query formulation but instead to assist them in formulating more effective queries in a shorter period of time. An interactive system, using our query formulation approach, could help the searcher to further expand and refine queries as they see fit, help in training purposes to allow students to 'get a feel' for how to search for literature without starting from scratch, or for researchers conducting rapid reviews: where time is a major factor and the accuracy of the review (with respect to finding every possible relevant study for the research question) is not as important as it is in traditional systematic review settings.

Overall, more effective query formulation has the potential to reduce the time spent screening and appraising literature, leading to more timely and cost-effective systematic reviews, leading to more up-to-date evidence based medicine and therefore more accurate diagnosis and decisions by clinical professionals.

Acknowledgements. Harrisen is the recipient of a CSIRO PhD Top Up Scholarship. Dr Guido Zuccon is the recipient of an Australian Research Council DECRA Research Fellowship (DE180101579) and a Google Faculty Award. This research is supported by the National Health and Medical Research Council Centre of Research Excellence in Informatics and E-Health (1032664).

REFERENCES

[1] Ion Androutsopoulos, Graeme D Ritchie, and Peter Thanisch. 1995. Natural language interfaces to databases–an introduction. *Natural language engineering* 1, 1 (1995), 29–81.
[2] Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 17.
[3] Justin Clark. 2013. Systematic Reviewing. In *Methods of Clinical Epidemiology*, Gail M. Williams Suhail A. R. Doi (Ed.). Springer.
[4] S José Closs, Dawn Dowding, Nick Allcock, Claire Hulme, John Keady, Elizabeth L Sampson, Michelle Briggs, Anne Corbett, Philip Esterhuizen, John Holmes, et al. 2016. Meta-review: methods. In *Towards improved decision support in the assessment and management of pain for people with dementia in hospital: a systematic meta-review and observational study*. Vol. 4. Health services and delivery research.
[5] Aaron M Cohen, William R Hersh, Kim Peterson, and Po-Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *JAMIA* 13, 2 (2006), 206–219.
[6] Gordon V Cormack and Maura R Grossman. 2016. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 1039–1048.
[7] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. 2014. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. ACM, 1819–1822.
[8] Julie Glanville, Sue Bayliss, Andrew Booth, Yenal Dundar, Hasina Fernandes, Nigel David Fleeman, Louise Foster, Cynthia Fraser, Anne Fry-Smith, Su Golder, et al. 2008. So many filters, so little time: the development of a search filter appraisal checklist. *Journal of the Medical Library Association: JMLA* 96 (2008).
[9] Sally Green and J Higgins. 2005. Cochrane handbook for systematic reviews of interventions.
[10] Diagnostic Test Accuracy Working Group et al. 2012. Handbook for DTA reviews. https://methods.cochrane.org/sdt/handbook-dta-reviews
[11] Elke Hausner, Charlotte Guddat, Tatjana Hermanns, Ulrike Lampert, and Siw Waffenschmidt. 2016. Prospective comparison of search strategies for systematic reviews: an objective approach yielded higher sensitivity than a conceptual one. *Journal of clinical epidemiology* 77 (2016), 118–124.
[12] Elke Hausner, Siw Waffenschmidt, Thomas Kaiser, and Michael Simon. 2012. Routine development of objectively derived search strategies. *Systematic reviews* 1, 1 (2012), 19.
[13] Jimmy, Guido Zuccon, and Bevan Koopman. 2018. Choices in Knowledge-Base Retrieval for Consumer Health Search. In *Advances in Information Retrieval*,

Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.). Springer International Publishing, Cham, 72–85.
[14] Evangelos Kanoulas, Rene Spijker, Dan Li, and Leif Azzopardi. 2018. CLEF 2018 Technology Assisted Reviews in Empirical Medicine Overview. In *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS.*
[15] Sarvnaz Karimi, Stefan Pohl, Falk Scholer, Lawrence Cavedon, and Justin Zobel. 2010. Boolean versus ranked querying for biomedical systematic reviews. *BMC MIDM* 10, 1 (2010), 1.
[16] Sarvnaz Karimi, Justin Zobel, Stefan Pohl, and Falk Scholer. 2009. The challenge of high recall in biomedical systematic search. In *Proceedings of the third international workshop on Data and text mining in bioinformatics.* ACM, 89–92.
[17] Youngho Kim, Jangwon Seo, and W Bruce Croft. 2011. Automatic boolean query suggestion for professional search. In *SIGIR'11.*
[18] Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1.* Association for Computational Linguistics, 423–430.
[19] Bevan Koopman, Guido Zuccon, Peter Bruza, Laurianne Sitbon, and Michael Lawley. 2012. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management.* ACM, 2439–2442.
[20] Robert Krovetz. 1993. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 191–202.
[21] Thorsten Kurz and Kilian Stoffel. 2002. Going beyond stemming: creating concept signatures of complex medical terms. *Knowledge-Based Systems* 15, 5-6 (2002).
[22] Matthew Lease, Gordon V CormackΔ, An T Nguyen, Thomas A TrikalinosV, and Byron C Wallace. 2016. Systematic Review is e-Discovery in Doctor's Clothing. In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR).*
[23] Grace E. Lee and Aixin Sun. 2018. Seed-driven Document Ranking for Systematic Reviews in Evidence-Based Medicine. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18).* ACM, New York, NY, USA, 455–464. https://doi.org/10.1145/3209978.3209994
[24] Mariska MG Leeflang, Jonathan J Deeks, Constantine Gatsonis, and Patrick MM Bossuyt. 2008. Systematic reviews of diagnostic test accuracy. *Annals of internal medicine* 149, 12 (2008), 889–897.
[25] Paige Martin, Didi Surian, Rabia Bashir, Florence T Bourgeois, and Adam G Dunn. 2019. Trial2rev: Combining machine learning and crowd-sourcing to create a shared space for updating systematic reviews. *JAMIA Open* (2019).
[26] David Martinez, Sarvnaz Karimi, Lawrence Cavedon, and Timothy Baldwin. 2008. Facilitating biomedical systematic reviews using ranked text retrieval and classification. In *ADCD'08.*
[27] Jessie McGowan and Margaret Sampson. 2005. Systematic reviews need systematic searchers (IRP). *Journal of the Medical Library Association* 93, 1 (2005), 74.
[28] Makoto Miwa, James Thomas, Alison O'Mara-Eves, and Sophia Ananiadou. 2014. Reducing systematic review workload through certainty-based screening. *JBI* 51 (2014), 242–253.
[29] Michael L Mortensen, Gaelen P Adam, Thomas A Trikalinos, Tim Kraska, and Byron C Wallace. 2017. An exploration of crowdsourcing citation screening for systematic reviews. *Research synthesis methods* 8, 3 (2017), 366–386.
[30] Rodolfo A Pazos R, Juan J González B, Marco A Aguirre L, José A Martínez F, and Héctor J Fraire H. 2013. Natural language interfaces to databases: an analysis of the state of the art. *Recent Advances on Hybrid Intelligent Systems* (2013), 463–480.
[31] Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics* 40, 3 (2007), 288–299.
[32] Ana-Maria Popescu, Alex Armanasu, Oren Etzioni, David Ko, and Alexander Yates. 2004. Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability. In *Proceedings of the 20th international conference on Computational Linguistics.*
[33] Martin F Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980).
[34] John Rathbone. 2017. *Automating systematic reviews.* Ph.D. Dissertation. Bond University.
[35] Scott Reeves, Ivan Koppel, Hugh Barr, Della Freeth, and Marilyn Hammick. 2002. Twelve tips for undertaking a systematic review. *Medical teacher* 24, 4 (2002).
[36] Eric Sayers. 2010. A General Introduction to the E-utilities. *Entrez Programming Utilities Help [Internet]. Bethesda: National Center for Biotechnology Information* (2010).
[37] Harrisen Scells, Daniel Locke, and Guido Zuccon. 2018. An Information Retrieval Experiment Framework for Domain Specific Applications. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.*
[38] Harrisen Scells and Guido Zuccon. 2018. Generating Better Queries for Systematic Reviews. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18).* ACM, New York, NY, USA, 475–484.
[39] Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2019. Automatic Boolean Query Refinement for Systematic Review Literature Search. In *Proceedings of The Web Conference (WebConf '19).* 1646–1656.
[40] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Shlomo Geva, and Leif Azzopardi. 2017. A Test Collection for Evaluating Retrieval of Studies for Inclusion in Systematic Reviews. In *The 40th International ACM SIGIR Conference on Research & Development in Information Retrieval.*
[41] Stefan Schulz, Martin Honeck, and Udo Hahn. 2001. Indexing medical WWW documents by morphemes. *Studies in health technology and informatics* 1 (2001), 266–270.
[42] Ian Shemilt, Nada Khan, Sophie Park, and James Thomas. 2016. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic reviews* 5, 1 (2016), 140.
[43] Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir.*
[44] James Thomas and Angela Harden. 2008. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC medical research methodology* 8, 1 (2008), 45.
[45] Mercedes Torres Torres and Clive E Adams. 2017. RevManHAL: towards automatic text generation in systematic reviews. *Systematic Reviews* 6, 1 (2017).
[46] Anton H van der Vegt, Guido Zuccon, and Bevan Koopman. 2019. Learning Inter-Sentence, Disorder-Centric, Biomedical Relationships from Medical Literature. In *AMIA Fall Symposium.*
[47] Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Brian Zhu, and Iain J Marshall. 2016. Extracting PICO sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research* 17, 132 (2016), 1–25.
[48] Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bio* 11, 1 (2010).
[49] Ye Zhang, Matthew Lease, and Byron C Wallace. 2017. Active Discriminative Text Representation Learning.. In *AAAI.* 3386–3392.
[50] Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103* (2017).