International Journal on Digital Libraries
https://doi.org/10.1007/s00799-023-00369-y

AgAsk: an agent to help answer farmer's questions from scientific documents

Bevan Koopman[2] · Ahmed Mourad[1] · Hang Li[1] · Anton van der Vegt[1] · Shengyao Zhuang[1] · Simon Gibson[2] · Yash Dang[1] · David Lawrence[3] · Guido Zuccon[1]

Received: 8 November 2022 / Revised: 9 May 2023 / Accepted: 19 May 2023

Abstract
Decisions in agriculture are increasingly data-driven. However, valuable agricultural knowledge is often locked away in free-text reports, manuals and journal articles. Specialised search systems are needed that can mine agricultural information to provide relevant answers to users' questions. This paper presents AgAsk—an agent able to answer natural language agriculture questions by mining scientific documents. We carefully survey and analyse farmers' information needs. On the basis of these needs, we release an information retrieval test collection comprising real questions, a large collection of scientific documents split in passages, and ground truth relevance assessments indicating which passages are relevant to each question. We implement and evaluate a number of information retrieval models to answer farmers questions, including two state-of-the-art neural ranking models. We show that neural rankers are highly effective at matching passages to questions in this context. Finally, we propose a deployment architecture for AgAsk that includes a client based on the Telegram messaging platform and retrieval model deployed on commodity hardware. The test collection we provide is intended to stimulate more research in methods to match natural language to answers in scientific documents. While the retrieval models were evaluated in the agriculture domain, they are generalisable and of interest to others working on similar problems. The test collection is available at: https://github.com/ielab/agvaluate.

Keywords Information retrieval · Professional search · Domain-specific search · Agriculture · Passage retrieval

1 Introduction

Twenty-first century agriculture is increasingly mechanised, data-driven and scientific-evidence based [2, 28, 35]. Even developing countries are seeing increasing digital disruption [11, 24].

A wealth of valuable resources and data could be used by agricultural users, but there are significant barriers in effectively accessing these resources. Much are locked away in large and heterogeneous datasets, research project reports, communications and scientific publications, meteorological and soil sample data, and external services and applications

Bevan Koopman and Ahmed Mourad have contributed equally to this work.

✉ Bevan Koopman
bevan.koopman@csiro.au

✉ Ahmed Mourad
a.mourad@uq.edu.au

Hang Li
hang.li@uq.edu.au

Anton van der Vegt
a.vandervegt@uq.edu.au

Shengyao Zhuang
s.zhuang@uq.edu.au

Simon Gibson
simon.gibson@csiro.au

Yash Dang
y.dang@uq.edu.au

David Lawrence
david.lawrence@daf.qld.gov.au

Guido Zuccon
g.zuccon@uq.edu.au

[1] The University of Queensland, Brisbane, Australia

[2] CSIRO, Brisbane, Australia

[3] Queensland Department of Agriculture and Fisheries, Brisbane, Australia

Published online: 19 June 2023

[18]. Some are structured data, while large amounts are still in natural language form. These natural language documents are not easily discoverable and synthesised. No federated service is in place that offers agricultural users a single entry-point to search this type of information. Thus, agricultural users are not able to put into practice valuable insights from such information.

On the other hand, digital connectivity is not the major barrier to accessing agricultural resources. Farmers now make use of handheld devices and digital services, Twitter being one popular platform for farmers to keep informed of the latest trends [19, 33].

The real barrier is how to effectively serve farmers complex, multi-faceted information needs. Scientific-like questions such as "What varieties of bread wheat are most resistant to crown rot?" are hard to answer automatically. Two problems make these questions difficult to answer:

- *Complex answer matching* Farmers may express their queries in ways that do not directly match relevant information. This complex information need also comes with many variations in how users would express their query/question. An automated system must handle such variations in a robust manner.
- *Focused answers* Farmers need easily digestible answers to their questions: presenting a 25 page scientific document will not do, both from a workload perspective and for farmers to recognised how it might relate to their query.

The above are common IR problems for which there are some existing solutions. For complex answer matching, neural models are currently state-of-the-art [9]. These methods do not rely on matching individual terms but instead rely on learned representations of word meaning. This breaks the dependence on specific terms used in queries and relevant passages and allows for 'semantic' matching; i.e. matching based on word meaning.

For producing focused answers, breaking documents into passages and ranking these against a user's query can provide the digestible answers users seek. Again, neural methods encode short passages of text into a representations that can be effectively ranked against a user's query—another short passage of text itself.

Contributions

This paper presents a framework that serves the information needs of agricultural users by:

- Analysing the information needs of real agricultural users, including the sources of information they use.
- Building a public dataset for evaluating search systems in a new and growing domain—search from scientific articles in general and in the agriculture domain in particular—which comprises a 86,846 document collection (further divided into 9,441,693 passages) carefully compiled by domain experts rather than web crawling or crowd sourcing. It also provides 210 rich, multi-faceted, real-world search topics comprising: (i) a natural language question; (ii) multiple keyword query variations; (iii) an expert-authored answer; and (iv) graded relevance assessment of passages.
- Providing a series of retrieval experiments with both baseline term-based retrieval models and state-of-the-art neural rankers.
- Providing an end-to-end system, AgAsk, that offers agricultural users a single entry-point to search this information.

These contributions touch on all aspects of the problem: from the needs of the users to the resources required to investigate the problem, the underlying machine learning model and a production search system.

2 Related work

While conversational agents have been proposed as a viable means to provide good answers to growers' questions [2, 28], a limited number of solutions have been proposed and explored.

A number of systems arose as a results of the release of a substantial dataset of farmer questions from the Kian Call Center (KCC).[1] KCC was a phone helpline service for farmers to consult with agriculture expert advisors about best practice and it was specifically tailored for the Indian market and agricultural context. Systems that used some portion of this dataset include AgriBot [13], FarmChat [11] and Krushi [21].

Agribot was developed to address growers information needs related to weather, market rates, plant protection and government funding opportunities. This conversational agent focused on the data of all Indian states collected over a 5-year period and relied on sentence embeddings (sent2vec [1]) and entity extraction to compute the similarity between a user question and a background of common question–answer pairs. Answers were sourced from an underlying agricultural knowledge base. Thus, unlike AgAsk, the knowledge base was not backed by a comprehensive collection of scientific evidence and required manual curation of a domain-specific knowledge base.

[1] https://data.gov.in/dataset-group-name/kisan-call-centre.

FarmChat was a speech-based conversational system that relied on decision rules and answers manually derived from the KCC data to identify answers on the IBM Watson APIs to perform intent identification and dialogue flow management. Much of the attention in FarmChat was on information access in a context of limited literacy and technology expertise in rural Ranchi, India, and on the information delivery modality (audio vs. audio+text). FarmChat focused only on one crop (potatoes); it did not leverage machine learning for extracting knowledge but instead relied on a manually built knowledge base. The drawback of this approach is that FarmChat did not scale easily and was difficult to maintain and link to information sources. While it helped to answer grower questions similar to AgAsk, it was highly tailored to one crop and one region (unlike AgAsk which is both crop and region agnostic).

Krushi was a conversational chatbot aimed to address growers information needs related to weather, plant protection, animal husbandry, market price, fertiliser use, government schemes and soil testing. This conversational agent focused on the data of the nine districts in Maharashtra, India, collected over a year. It utilised the RASA X conversational AI system, involving intent identification followed by response retrieval. It was made accessible to farmers via WhatsApp.

Besides Indian resources which facilitate access to agricultural data that support farmers in rural areas, other resources have been developed in other countries. A user study collected 1000 Taiwanese conversations from interviews between investigators and farmers discussing specific topics and was developed to address sales, logistics and plants [7]. These data were utilised to train a LSTM sequence-to-sequence conversational model which relied on word embeddings to generate an answer to the input question.

Another resource developed a crop protection information system to support farmers in rural areas of Tanzania where it is hard for government agricultural officers to visit in a timely manner during seasonal diseases outbreaks [31]. A collection of 2100 Swahili queries were gathered from face-to-face interviews with 100 farmers. The authors analysed farmers' preferred method of expressing their information needs (keyword queries or natural language questions, and via SMS or the Web). They showed that there is a significant association between the age of farmers and their preferred method for expressing their information need, with the majority of young farmers (< 40) preferring short and simple SMS queries while old farmers preferring natural language questions.

While all the aforementioned resources help advance the digitisation of agriculture, they have few key limitations: (1) they are limited to either specific regions or crops so do not generalise; (2) the question-answer pairs are not grounded to the source of information (e.g. a research article); and (3) scalability is hampered by manual curation of the data rather than leveraging machine learning for extracting knowledge [12, 31].

Despite the increasing availability of rich data resources for farmers to draw on, there is a dearth of search-based systems that can bring this data together to answer a farmer's query. The few examples of such search-based agents in the agricultural sector, although limited in scope, showed promise and indicate that a larger effort in this area would be fruitful.

AgAsk addresses many of the aforementioned limitations by (1) being crop, region and question type agnostic; (2) being backed by a large collection of rigorous scientific information; (3) using automated methods to extract information, making the system scalable and avoiding manual curation; and (4) using state-of-the-art neural ranking models to match users questions to relevant passages (not documents) in the collection.

3 Information needs of users in agriculture

Users in agriculture can be broadly categorised into three types: growers (farmers), agronomists and specialists. The latter two are the experts that provide support to the farmers (either through paid consultations or sponsored by the government) and communicate to them the outcomes of recent research. The information needs of three user types overlap to a large degree, with some specific needs for each.

In this section, we first survey the literature on information needs of these users. The available literature pertains mostly to growers. Then, we conduct an online survey to gather the information needs of agronomists and specialists. The learning and materials detailed in this section will feed into the creation of an agricultural-specific test collection designed to evaluate search systems in this context; this is presented in Sect. 4.

3.1 Types of information needs from the literature

From the literature we summarise the specific information needs of growers. We constrain our analysis to those farmers involved in crop production (i.e. growers) and exclude animal production. While much of the concepts outlined here are relevant to both, animal production includes substantial veterinary content, excluded for the benefit of brevity.

From the literature, [6, 11, 28] some key categories of information needs were identified and are outlined below.

3.1.1 Crop protection

A significant number of grower's questions relate to protecting their crop from diseases or pests, whether for future

prevention or because of an existing outbreak. In the latter cases, farmers often describe crop diseases via visible symptoms (e.g. "brown spots on the leaves") in order to first identify the diseases and second determine the best course of treatment (e.g. what fungicide to use, including dosage and application instructions). Similarly, they may describe pest species (e.g. "2 cm black and yellow snail") to determine the relevant pesticide to use. Many queries relate to identifying and eradicating weeds [6]. For all these queries, it is important to point out that the grower's query typically does not contain keywords that match the relevant answer (e.g. the actual pest species name); instead, this needs to be inferred from the description of the symptoms / problems.

3.1.2 Best practices

Growers are constantly on the lookout for how they can increase the quantity or quality of their yield as well as reduce their costs or wastage, consequently increasing profitability. Agriculture is constantly evolving with new products and practices; many growers feel that keeping abreast of current best practice is critical [28]. While growers will ask specific questions on a topic when they require information, they also seek out recommendation services that "push" relevant information. For example, the use of Twitter is one common way of keeping abreast of trends [19, 33].

Best practices can cover the full spectrum of important topics in agriculture (agroecology, water management, etc.).

3.1.3 Unbiased product recommendations

Growers rely heavily on many agriculture products to run their farms. These can constitute a significant expense and as such they would like reliable and trustworthy product recommendations. Recommendations for different types of fertiliser, seed and crop variants and herbicide or pesticide are some commonly sought examples [11].

3.1.4 Markets and weather/climate

While the market and weather are factors outside grower's control, they will certainly wish to understand and adapt their practices to changes in both. Because a farm is a business producing agricultural products, it has the same requirements of access to and understanding of markets that all businesses have. Growers would like to understand and adapt to the market in which they operate [6]. This includes understanding of current and projected prices on products they sell as well as costs of products and services they consume.

Growers would like to take into account the past, current and future weather and climate. Planting, for example, is often tied specifically to periods of rainfall. Similarly, pest outbreaks often relate to weather and climate. Thus, growers would want any information returned to be tailored to the recent weather. Similarly, upcoming weather impacts grower's decisions so information should be tailored to weather forecasts. Longer-term climate information—both historic and projected—is also important to growers and needs to inform what information is presented to them.

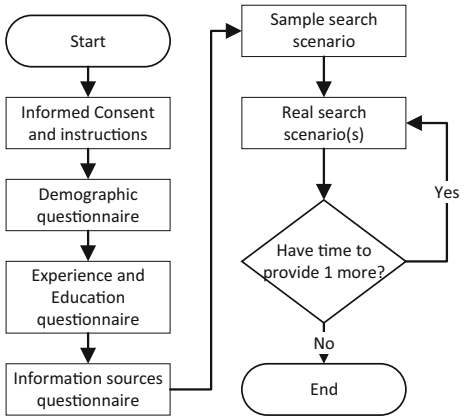

Fig. 1 Flowchart showing the steps involved in our online survey of to study the information needs of agricultural users. The 'sample search scenario' given to every participants was to find an answer to "How much nitrogen fertiliser will I need to put on my crop this year, following the drought?" using any method they prefer. The 'real search scenario' was an actual question, provided by the participant, that they had faced in the last 12 months

3.2 A survey to better understand expert users

To gain an accurate understanding of users' information needs and the resources they use to answer these, we conducted an online survey that targeted farmers, agronomists and agricultural specialists in Australia.[2]

3.2.1 Survey methodology

Fig. 1 depicts the flowchart of the survey. After consent, a questionnaire solicited information about the users' demography, prior experience and education, and previous sources of information they used to find answers to their question.

After this, participants were presented with a sample search scenario in the form of the question "How much nitrogen fertiliser will I need to put on my crop this year, following the drought?" This was a question identified by one expert as being (1) commonly sought after and (2) without an obvious answer. Participants had to then find an answer to this question through whatever means they felt best. This sample search scenario was done by all participants. The sample search scenario was used for two purposes: (1) to understand the strategies different users adopt (including which sources

[2] Ethics was granted by the University of Queensland for application #2020000826.

Table 1 Survey participants were asked to prove a real search scenario that they had had in the last 12 months. For each scenario, there were asked the following questions

Q1.	How important is answering the question to farm or crop success?
Q2.	How frequently does it arise?
Q3.	How urgent is it that you get an answer in a timely manner after it has arisen?
Q4.	Select the top 5 sources of information you would go to in order to help you derive an answer to the sample question
Q5.	Assuming that you had to search on the GRDC website or search on Google for an answer, please type in at least 3 different search queries that you would use to find information to help you
Q6.	Write down at least 3 elements of the answer that you would like to see as part of the complete answer to the question
Q7.	How much information would you like to receive in the answer
Q8.	How might the answer be contextualised much more to your situation? List at least 3 additional specific information items
Q9.	Write down a short summary answer, in 1--2 lines, if you know it
Q10.	If this is a question you have previously sought an answer to in real life, how successfully was the answer provided?

they look for) to identify answers for a controlled information need, (2) to provide an example search scenario that would familiarise them with the proceeding real search scenario task.

After completing the sample search scenario, participants were asked to provide two or more real search scenario of their own. They were asked for a real scenario that they might have had in the past 12 months and had to seek an answer to; i.e. questions that they could not answer with their own knowledge.

For each search scenario, participants were asked the questions shown in Table 1.

Recruitment of participants was done through the professional network of a contact at the Queensland Department of Agriculture and Fisheries. Participants were not paid.

3.2.2 Survey results and analysis

In total, 16 participants completed the survey. While the number of participants is not representative, we share some of the valuable insights that influenced our decisions for building the test collection described in Sect. 4.

Table 2 shows some statistics from the survey. Participants were divided among grain crop specialists (9) and agronomists (7). The majority had at least 10 years of experience and a bachelor degree. They tended to search for information that had significant bearing on farm or crop success with 70% of the search scenarios either essential or extremely important. They also tended to be patient with their search with 65% accepting to obtain an answer within days or up to a week. This suggests that answering agricultural users' information needs might be a slow search scenario, where you trade-off speed in favour of a high quality search experience [30].

Table 2 Statistics of the information needs survey

Role	16
Grain grower	0
Grain crop specialist	9
Agronomist (farm consultant)	7
Years of experience	16
10 years or more	10
Between 5 and 9 years	4
Between 1 and 4 years	1
Less than 1 year	1
Education	16
Doctoral degree	2
Master degree	1
Bachelor degree	10
Diploma	2
Vocational certificate	1
Perceived importance of search scenarios	64
Essential	20 (31.2%)
Very important	25 (39.1%)
Moderately important	16 (25.0%)
Somewhat important	1 (1.6%)
Not important	2 (3.1%)
Urgency of obtaining an answer	64
Extremely urgent (day)	7 (10.9%)
Very urgent (days)	22 (34.4%)
Urgent (week)	20 (31.2%)
Somewhat urgent (weeks)	9 (14.1%)
Not urgent	6 (10.9%)

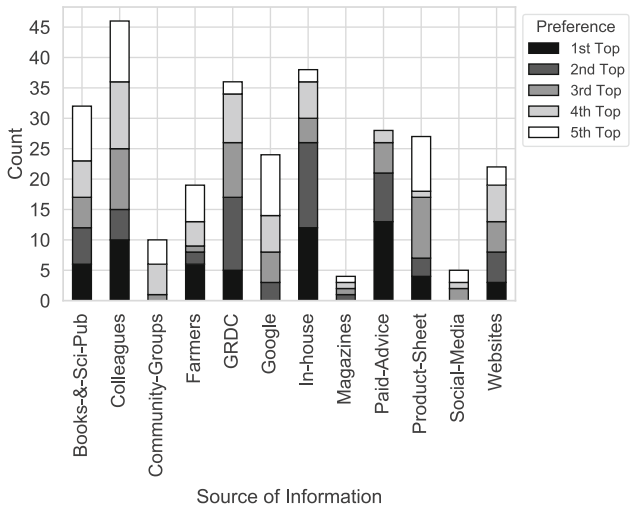

Fig. 2 Preference for sources of information (frequency)

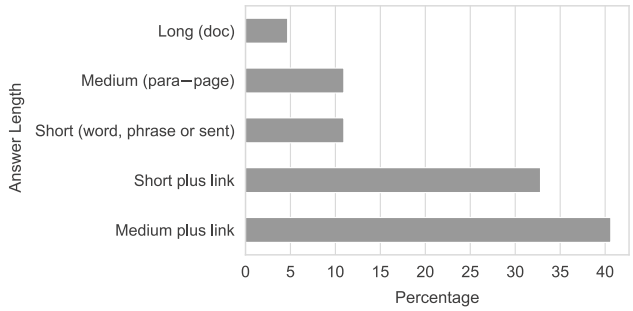

Fig. 3 Answer length (percentage)

Figure 2 depicts the preference for sources of information. We asked the users to select the top 5 out of 12 sources of information they would go to in order to help them find an answer to their questions. Agricultural experts tend to seek information from a wide range of sources with different levels of preference. They tend to trust more in-house reports (generated within the same organisation), colleagues, scientific publications, and paid advice. This is contrary to previous research that suggests Twitter is a popular platform for farmers to keep informed of the latest trends [19, 33]. This finding in specific, along with expert feedback from agronomist researchers in academia, government department and a leading research organisation, has influenced the selection of information sources for the documents collection.

Figure 3 shows the preference for the amount of information users would like to receive in the answer for each search scenario (Q7 in Table 1). Agricultural experts tend to prefer either short (single word, phrase or sentence) or medium (between a paragraph and a page) length answers with links to the evidence for further reading if required. Figure 4 demonstrates how successfully was an answer provided for each search scenario (Q10 in Table 1). Agriculture experts were very successful only 15% of the time. More than 65% of the

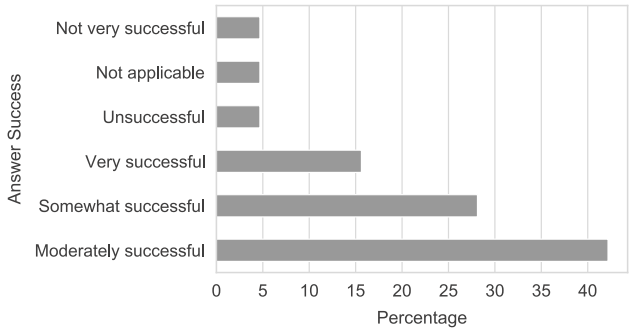

Fig. 4 Answer success (percentage)

answers provided were somewhat (a partial answer was provided) or moderately (a good answer was provided, although they would have preferred more information) successful.

While the analysis of the survey is limited by the number of participants, there were some key insights that were gained. It showed that users seek information from different sources—this influenced what sources we collected for AgAsk. The survey told us that users like moderate length answers with the option for more information—this informed the decision to show passages in the search results with the URL to the source document. The survey indicated that users generally have a pressing need to obtain answers to their questions—this made us understand there was demand for an interactive, question-answering system like AgAsk. While the survey did not provide extensive data for analysis, it did provide some valuable insights into users in this domain.

4 Development of an open dataset for empirical evaluation

The analysis of information needs in the previous section provides insights into how users go about looking for information in this domain. Lessons from the survey were then used to create an actual dataset for developing and evaluating search systems in this domain. We used the dataset for the development of the AgAsk system, but it is also a general resource available to others. In information retrieval, such resources are called "test collections" and are a key resource underpinning information retrieval research in a new area [36]. This section is dedicated to describing how the test collection was created and an analysis of its characteristics.

4.1 Methods to create the test collection

Creating a test collection involved three main steps: (1) obtaining a source of documents that users are interested in searching; (2) creation of a set of realistic questions that users would ask; (3) performing human relevance assessment that

judges the relevance of each question against documents in the collection. We detail each below.

4.1.1 Documents and passages

Two sources of agricultural information were obtained as part of the collection:

- **Industry reports** 4003 agricultural reports from the Grains Development Research Corporation and the State Departments of Agriculture in Australia.
- **Scientific articles** 82,843 scientific journal and conference articles from 33 agricultural journals.[3]

These selected reports and journal articles were considered relevant to the grains industry and focused on crop agronomy and soils. The targeted subject matter related to the growth and management of grains crops including cereals (e.g. wheat, barley, and sorghum), legumes (e.g. chickpea, soybean, mungbean), and oilseeds (e.g. canola), and the management of the soils on which these crops are grown. Topics covered included recommendations and research relevant to the management of individual crops through varietals selection, sowing times, planting rates and row spacing, etc.; whole farming system performance, crop sequencing and fallow management practices; fertiliser management; and the identification and management of pest and diseases that affected the grains industry. Both these sources came in the form of PDF documents.

The industry reports we collated are made publicly available.[4] The journal articles, instead, come from subscription journals so cannot be redistributed; however, we provide crawler scripts that can be used to download the full text using an institutional or paid subscription to these journals.

Once full-text PDFs were obtained, they were converted from PDF to JSON using Apache Tika. (Code for this is provided in the collection repository so that the processed collection can be fully reproduced, along with the pre-processed JSON files for the reports.) From here, the documents were further split into passages of three sentences. (The Spacy sentenciser was used to derive sentence boundaries and code is provided for this.) From the 86,846 documents, 9,441,693 passages were produced.

4.1.2 Creating questions/queries

Originally, questions were intended to come from the real search scenarios of the survey in Sect. 3. However, the small number of participants meant an alternative source for questions was needed. For this, we followed a process called known-item retrieval [23]. The process involved two human assessors (both agricultural scientists) creating questions via the following:

1. We randomly sampled a document from the collection and showed it to the assessor. If the document was not suitable for generating a reasonable question, the assessor could request another random document.
2. On reading the document, the assessor was asked "What question does this document help answer?". The question they provided became the natural language question.
3. They were asked to provide 3 or more (unlimited) ad-hoc, keyword search queries that correspond to the question.
4. They were asked to author an answer, in their own words, to that question.
5. They were asked to select and paste the relevant portion of the document that helped answer the question.
6. They were presented with a list of other passages from that document and asked to assess these as either relevant, marginal or non-relevant.

The result of one iteration of the above process is a single-question "topic" containing question, keyword queries, answer, relevant snippets and labelled passages. A sample topic created using the above method is shown in Fig. 5. The process was repeated by the human assessors to create 210 topics from 165 documents. (Multiple, different topics could sometimes be derived from a single document.)

Topics were divided into training and test sets. The 50 topics with the most relevance assessments formed the test set and the remaining 160 topics formed the training set. (Other splits can be done as desired; ours was purely done for our later experiments).

4.1.3 Human judging

Each of the 210 topics created only contained a handful of passages manually judged for relevance—insufficient for a rigorous evaluation. So we set out to perform more thorough manual assessment of passages. For each question topic, we aimed to manually assess passages as "Relevant", "Marginally Relevant" and "Not Relevant" to that topic. Assessing all 9+ million passages for each topic was clearly not possible, so we used the standard information retrieval process of forming a judgment pool [27]. This is typically done by running the questions through a few search systems and then judging the top ranked results that each system returns.

We considered two state-of-the-art neural ranking systems, which we also used then for experimentation: monoBERT and TILDEv2. (These models are detailed later in Sect. 5.1.)

[3] Agricultural scientists and authors Y.Dang and D.Lawrence compiled a list of relevant journals.

[4] https://doi.org/10.48610/fa4684b.

Question:	What type of herbicides are effective against sowthistle?
Keyword queries:	sowthistle herbicide mixing Balance A Group D Group K sowthistle broadleaf active herbicides mixing
Assessor Authored Answer	The addition of Balance to either Group D or Group K herbicides can provide good control of sowthistle. The addition of Flame, Group D, Balance or Group K to broadleaf active herbicides (Group C and Valour) are also effective.
Relevant Passages:	20171829-40 (Relevant): In the trials reported here the addition of Balance to either Group D or Group K has provided good control of sowthistle, when these same products applied alone are not providing acceptable control. ... 20171829-39 (Marginal): The trials reported here demonstrated that these products can perform quite poorly on the broadleaf weed sowthistle, when applied alone. The most effective products for sowthistle in these trials were Valor ... ...

Fig. 5 Sample topic from the test collection. Each topic contains a question, a number of ad-hoc queries, an answer authored by assessors and a list of relevant passages graded with "Relevant", "Marginally Relevant" and "Not Relevant"

These models represent state-of-the-art passage retrieval systems: monoBERT being the most effective empirically but having prohibitively slow query latency for interactive systems; TILDEv2 being the most effective method that still has good query latency for interactive systems.

Results for all 210 topics were produced with monoBERT and TILDEv2. These results were fused using reciprocal rank fusion to produce the final pool for human assessment [15, 16].

Relevance assessment was conducted by authors D.La wrence and Y.Dang, both agricultural scientists. We developed a custom software tool called Agotator to support accurate and rapid relevance assessment.[5] A screenshot is shown in Fig. 6.

As seen from the screenshot, users were presented with the topic question, a list of passages for judging, along with a link to the PDF source document from which the passage was extracted. Grades of relevance were: relevant, marginal and non-relevant. The criterion for relevance given to assessors was: "does the passage help to answer the question", where 'relevant' meant that the passage contained the answer, 'marginal' meant the passage contained some part but not the whole answer, and 'non-relevant' meant the passage contained no useful information.

For the topics from the training set, assessors judged the top 10 passages. For the topics from the test set, assessors judged the top 20 passages; if no relevant passage was found in the top 20 then they continued down the ranking until a relevant passage was found or rank 100 was reached. This procedure ensured that test topic was more thoroughly judged and was highly likely to contain relevant passages.

4.2 Characteristics of the test collection

Table 3 provides statistics for different parts of the test collection. Topics in the collection were multi-faceted, containing a natural language question, a number of keyword queries, a human authored answer, and relevance passages; a sample topic is shown in Fig. 5.

As seen, the test collection supports query variations by having multiple keyword queries for each topic. Figure 7 shows a histogram of the number of keyword queries for each topic. Most topics contain three queries (mean = 3, SD = 0.92), as per the instructions to assessors to provide at least three. Query length in number of words is shown in Fig. 8. As to be expected, natural language questions were both longer and more varied in length. Most keyword queries were between 3 and 4 words long.

Figure 9 shows the breakdown of grades of relevance for the topics in training and test sets. Recall that for the test set, assessors judged to rank 20, stopping there if at least one relevant passage was found, otherwise continuing down the ranking until a relevant passage was found or rank 100 reached. As seen from the plot, no relevant passages were found for one topic. We opted to keep this topic in the collection because it was from the known-item retrieval set, which means there was at least one relevant passage, but that had not been retrieved by any of our models in the pool.

[5] We plan to open source Agotator in a future work.

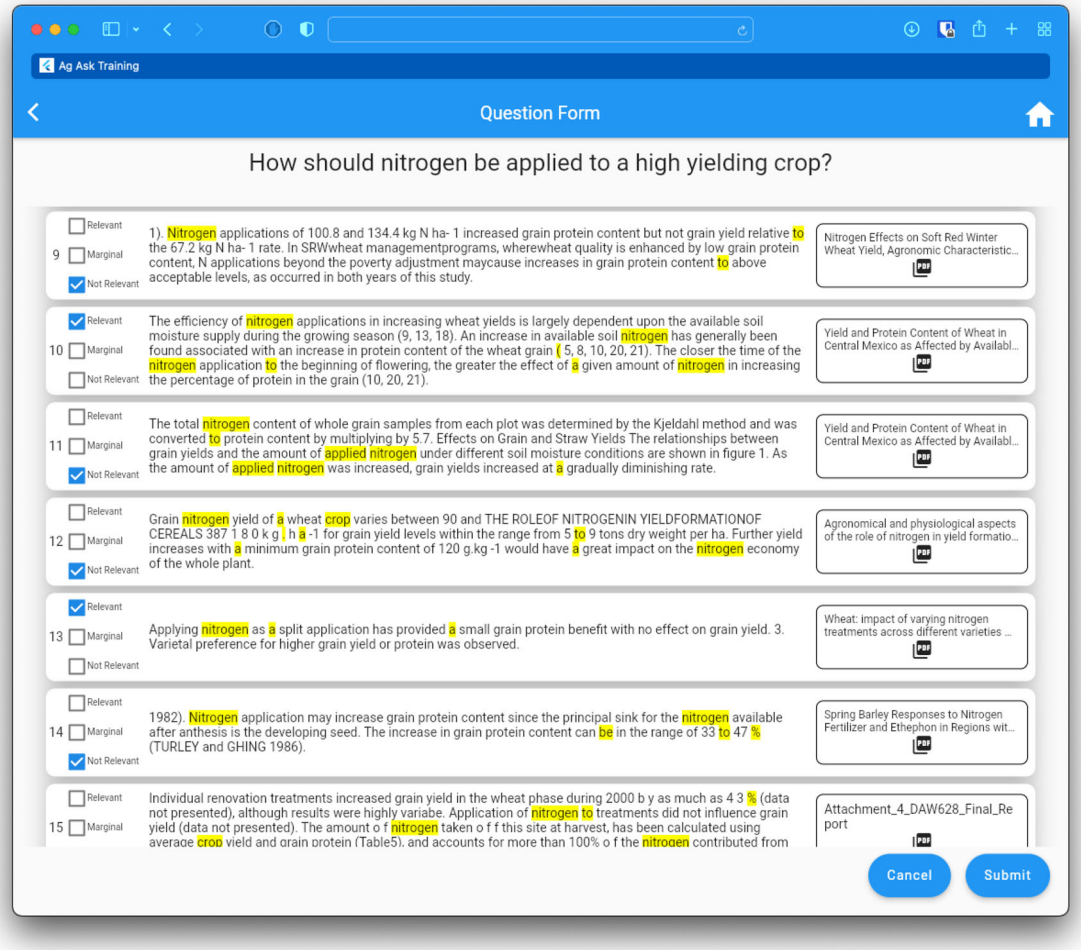

Fig. 6 Relevance assessment of passages using the Agotator tool. The yellow highlight indicates shared terms from the question and was provided to aid with relevance assessment

Table 3 Statistics of the test collection we compiled

Topics	210
Train	160
Test	50
Judged Passages	3948
Non-relevant	1244 (32%)
Marginal relevant	852 (22%)
Relevant	1852 (48%)
Documents	86,846
Reports	4003
Journal articles	82,843
Passages	9,441,693

5 Passage retrieval

Two main experiments were conducted: (1) understanding the effectiveness of a selection of retrieval models on this collection; (2) understanding how query variations impact effectiveness.

5.1 Retrieval methods

We implemented the following retrieval methods:

- **BM25:** Vanilla BM25 baseline to understand how a simple term-based retrieval performs.
- **BM25-RM3:** a BM25 baseline with pseudo-relevance feedback using RM3.
- **monoBERT**: a cross-encoder neural method involving a first-stage BM25 initial retrieval of 1000 documents, followed by a fine-tuned monoBERT reranker [22]. We used a monoBERT model pre-trained on the MSMARCO dataset and then fine-tuned on the 160 training topics.
- **TILDEv2** is a neural reranker that utilises document expansion at indexing time to avoid the need for neural encoding of documents at query time [38]. It involved

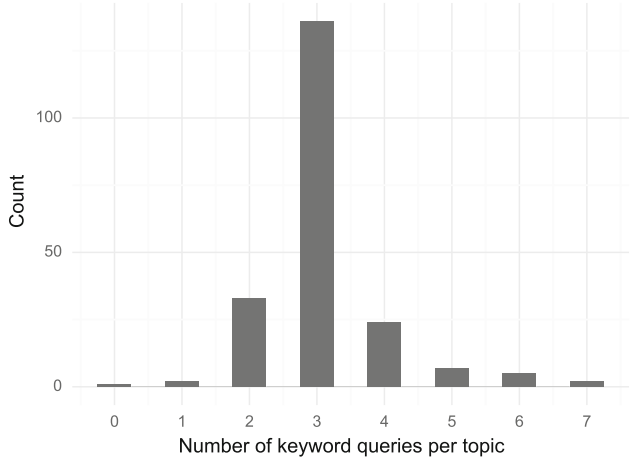

Fig. 7 Histogram showing the number of keyword queries for each topic. Mean = 3, SD = 0.92

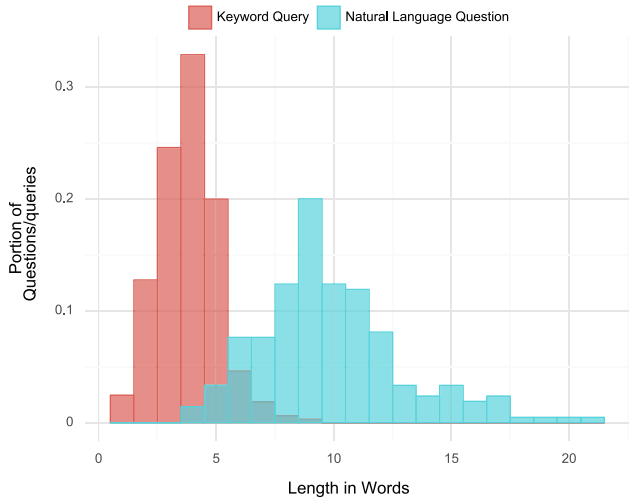

Fig. 8 Query length in number of words for natural language questions and keyword queries. Mean length for questions = 9.8 words and keywords = 3.8

a first-stage BM25 retrieval of 1000 documents, followed by a fine-tuned TILDEv2 reranker. TILDEv2 was added as a computationally efficient—yet still effective—model that might be deployed in a live search system. This model was also fine-tuned on the 160 training topics.

To make use of the multi-faceted topics provided in the collection, we ran the above models using both the natural language questions and keyword query versions of the topic. This aimed to uncover some insights into how query variation impact effectiveness.

5.2 Results

The effectiveness of the above models is shown in Fig. 10.

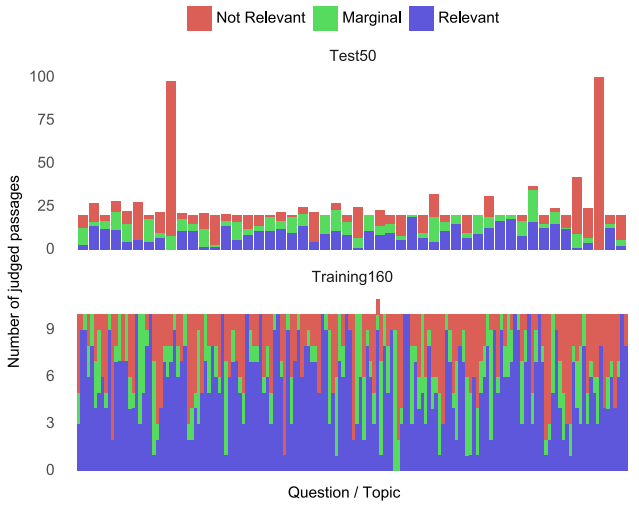

Fig. 9 Breakdown of relevant and non-relevant passages for each question in the training and test sets

5.2.1 Term-based vs. neural model effectiveness

There was a large difference in effectiveness between the term-based BM25 model and the neural rankers on this collection: monoBERT and TILDEv2 models were far more effective than BM25 (t-test, $p < 0.01$ for nDCG@5). However, it's worth noting that for measures like Success@100, BM25 was highly effective (no statistically significant difference between BM25 and neural models). This meant that BM25 *retrieved* the relevant passages, but was not effective at *ranking* them (low effectiveness for measures that consider top ranked results; e.g. NDCG@5). This tells us that using a BM25 for initial retrieval was reasonable, if it was followed by a high-precision reranker.

5.2.2 Natural language vs. keyword queries

Using natural language questions was more effective than keyword queries in most cases. (This is somewhat contrary to previous research that has shown verbose queries are less effective [5].) The neural models, in particular, were suited to questions rather than queries. The benefit of using questions is seen in early precision and not recall; i.e. improvements were seen in measures such as NDCG@5 and reciprocal rank that measure early precision rather than success@100 that measure recall.

The popular technique of pseudo-relevance feedback on top of BM25 (i.e. RM3) actually reduced effectiveness for keyword queries. However, pseudo-relevance feedback was effective when applied to natural language questions.

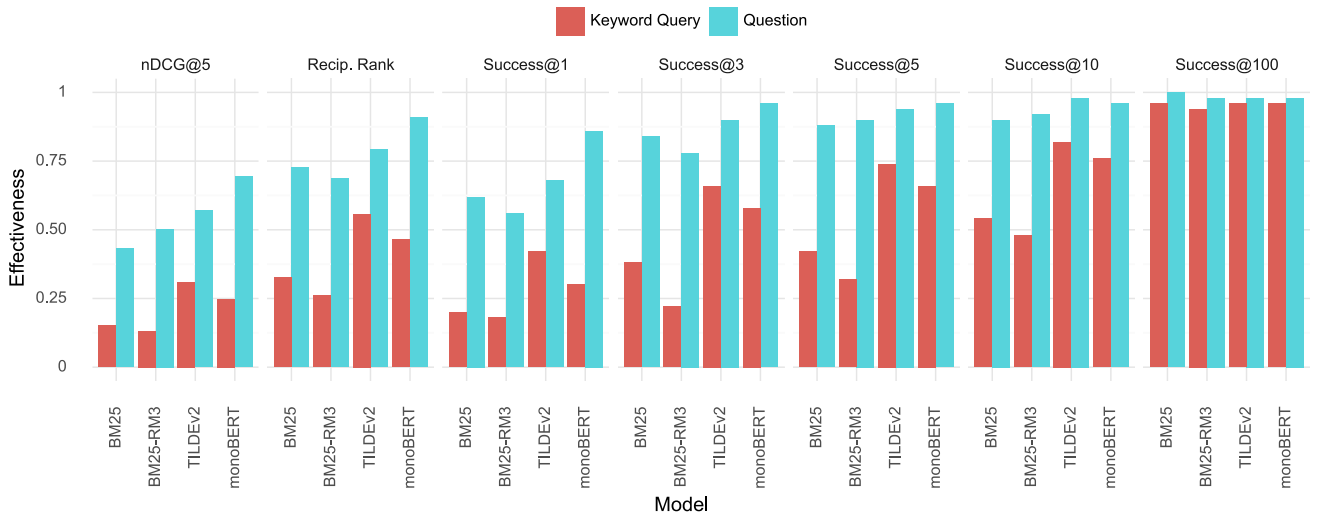

Fig. 10 Retrieval effectiveness of different models. Both natural language questions and keyword query topic types were evaluated

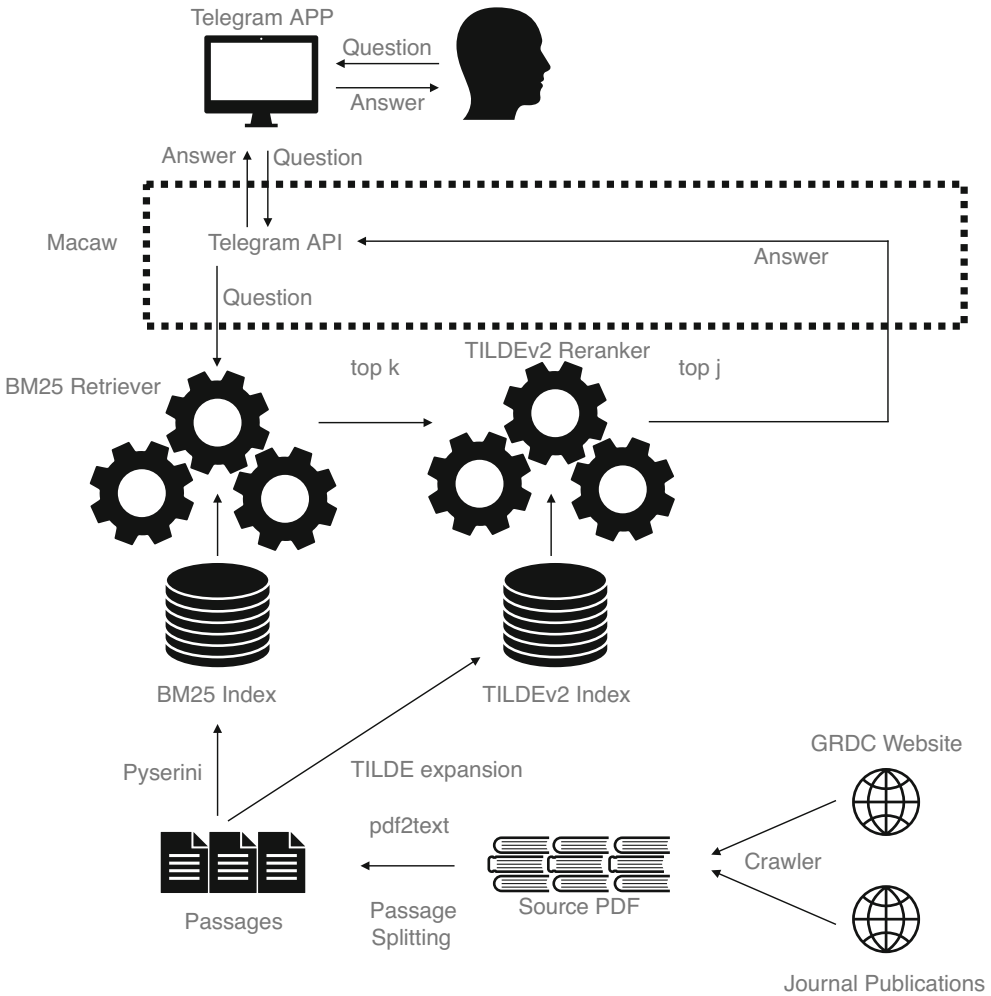

Fig. 11 Overall architecture of AgAsk

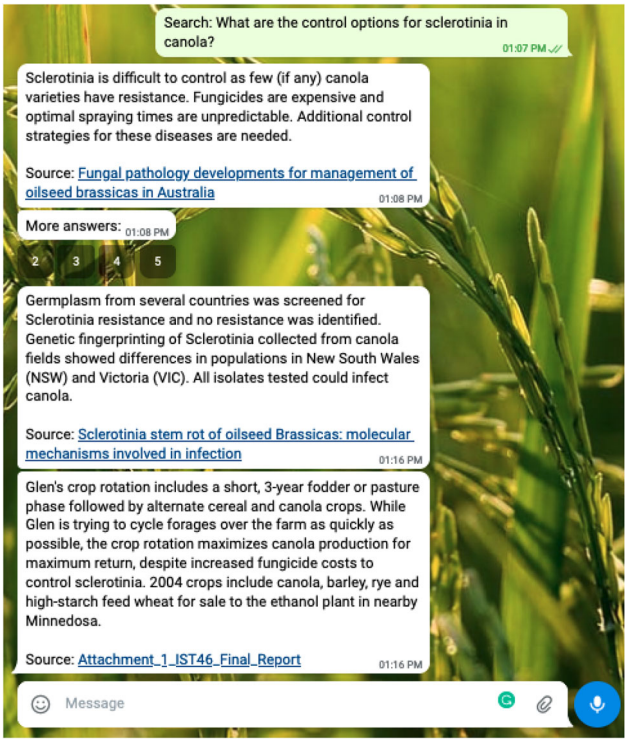

Fig. 12 A screenshot showing a AgAsk in use. Top is the user's question, along with best match answer passage. Buttons under "More answers:" allow the user to see the next four ranked passages, two of which are displayed. Each document from which the passage was extracted is shown as a hyperlink next to "Source:"

6 End to end integrated solution: AgAsk

In this section, we describe our single entry-point system for agricultural users to help them search for information, dubbed AgAsk. AgAsk can be deployed as a conversational agent, or a traditional search engine. Figure 11 provides the overall architecture of AgAsk in its deployment as a conversational agent. We utilise the Telegram messaging platform to handle messaging. Users submit their question via the Telegram 'AgAsk' bot. Overall conversation management is handled by Macaw [37], an open-source framework for building conversational search systems. Macaw passes the query to our custom retrieval pipeline, comprising of a first stage BM25 retriever and the neural TILDEv2 re-ranker [38]. Retrieved passages are then sent back to Macaw, which is responsible for serving it back to the grower via Telegram.

6.1 Client and user interface

An example AgAsk session in Telegram is shown in the example screenshot of Fig. 12. Telegram was chosen because it provides a simple API and Telegram clients are available for every major platform and device. The grower can pose a natural language question and AgAsk will respond with a generated answer.

A demonstration video of AgAsk is available at https://ielab.io/projects/agask.html. The clarifying questions are currently manually inserted to demonstrate what a fully interactive system might look like. We are in the early stages of deploying in production such a mixed-initiative conversational system.

We also log all user interactions including clicks, likes and emojis. This provides a source of relevance feedback information that may be used in future feedback mechanisms or online learning to rank.

6.2 Conversation management with Macaw

AgAsk employs the Macaw conversational information seeking framework [37], as it provides a convenient way of building an entire pipeline from scratch. The Macaw framework consists of several modules, including intent identification, co-reference resolution, query generation, retrieval model, and result generation. Currently, we have disabled the intent identification, co-reference resolution, query generation, file IO, and standard command line IO modules. We have instead instantiated our own retrieval and result generation modules, as detailed above, while we are in the process of deploying in production relevant modules for intent identification, relevance feedback, and question clarification.

6.3 Choices in retrieval model

The monoBERT reranker was the best performing model (see from Sect. 5). If you consider a live question-answering system that might provide three possible answers to a user's question (e.g. in a conversational or mobile setting) then success@3 would be the measure to consider. In this setting monoBERT provided a success@3 of 0.96: 48/50 topics had a relevant passage in the top 3 results. We posit this would make for a highly effective real system if the results generalise beyond the test topics in our collection.

While monoBERT was highly effective, it was computationally expensive. Query latency would make it prohibitive for real users in an online passage retrieval setting; or specialist GPU and parallel hardware might be required. TILDEv2, while less effective, was far more efficient and could be deployed in production on commodity CPU-based hardware (although document expansion and indexing were best done using a GPU).

Figure 13 depicts the effectiveness-efficiency tradeoff for different retrieval models for AgAsk. It suggests that achieving more effectiveness requires more query latency. This is particularly evident when comparing monoBERT to either BM25 or TILDEv2. monoBERT achieves a higher NDCG@5 with a considerable trade-off in latency. On the

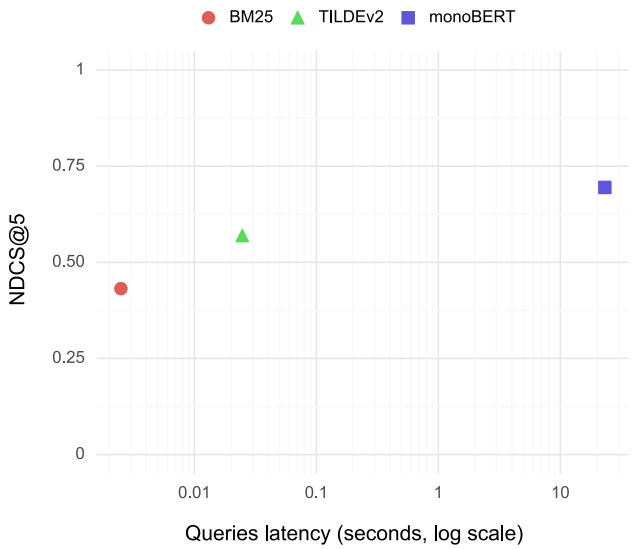

Fig. 13 The effectiveness–efficiency tradeoff for different retrieval models for AgAsk. While monoBERT is the most effective, it cannot serve queries to a user in a timely manner. TILDEv2 offers a far more efficient model with only a small reduction in effectiveness; hence, TILDEv2 was chosen as the underlying model for AgAsk

other hand, TILDEv2 strikes a great balance between effectiveness and query latency. Hence, we employ TILDEv2 in AgAsk. A further advantage of using TILDEv2 is that it does not need a dedicated GPU-based server to be used in production, as monoBERT does instead, as TILDEv2 runs entirely on CPU for its inference stage.

7 Future work

7.1 Further research using the test collection

The test collection detailed in Sect. 4 is a standalone resource that can be used, independent of the AgAsk system, in the development and evaluation of search systems for the agricultural sector.

Passages vs. documents The collection contains both full documents and sub-document passages. This allows other researchers to investigate differences in effectiveness between passage and document retrieval [8, 14, 17].

Query variation Topics in the collection are multi-faceted: they contain a natural language question and multiple keyword queries. Query variations have a large impact on retrieval effectiveness [20] and the study of query variation is an active research area [3]. The test collection provides a query variation resource. The fact that it contains both natural language questions and keyword queries means that these two different representations can be analysed by others.

Answer generation For each topic, assessors authored an answer to the topic question in their own words. (The sample topic from Fig. 5 shows this.) Note that these answers may differ in vocabulary or substance from relevant passages from the document. They represent the assessors expression of what the document contains. Using this, the collection could be used to develop and evaluate answer generation methods that, for example, take a set of retrieved passages to derive a natural language answer to the question [10]. The answers could also be analysed to understand how similar or different the language used in answers is to that of relevant passages. We have already began work on answer generation, in particular using large language models such as ChatGPT for this task.

Scientific document extraction The reports and journal articles used in the collection were processed using a basic PDF extraction method that divided the document into three sentence passages. This method did not account for the structure of the document—paragraphs, figures, tables, sections, etc. Using the collection, one could investigate many different information extraction methods for scientific articles [4], and the impact that these have on retrieval, answer generation or any other downstream tasks. This is a immediate piece of work that could improve passage quality.

Domain specific/expert search : Previous research has demonstrated the value of and the need for domain-specific test collections to evaluate the effectiveness of open-domain information retrieval models [26, 29, 32, 34]. However, there is no search datasets for agriculture. Our test collection represents a domain-specific, expert scenario. Models that work in the open domain may not translate to this expert domain. The test collection provides a resource to test this and possibly develop new models suited to this domain.

7.2 The case for contextualisation

AgAsk matches a query to a passage without taking into account characteristics of the user—there is no personalisation or contextualisation. Through our analysis of the information needs of users in agriculture, we identified that certain characteristics of a user have a strong bearing on their information need and impact on what they would consider relevant answers to their questions. We detail some below. We further note this need for contextualisation and adaptation to the different practices of the individual users within the same professional domain is a common characteristic across other professional search tasks [25].

Weather and climate Information should be tailored to recent weather, forecasted and longer term climatic predictions (e.g. if the farmer is located in a drought predicted area

then recommendations for drought resistant crops would be important).

Location The grower's region strongly informs their information need. The growing conditions, access to markets, infrastructure (e.g. rail or irrigation networks), historical crop yields and many other factors can be inferred from location. Thus, growers would like information that is location-aware.

Markets Contextualisation to the specific market that the grower operates in, including price, trends and changing customer demands/preferences.

Literacy/interpretably Evidence-based agriculture involves making decisions based on scientific evidence and sources. While growers may recognise the value of this, they do not necessarily want to delve into detailed scientific information, or have the expertise to do so. Instead, they would like outcomes of the scientific literature to be provided to them in an understandable, concise and digestible form. Furthermore, grower's expertise varies considerably—some may have detailed technical expertise in certain areas and thus would like to see associated technical details; others may have no technical expertise in the area and require a lay overview. Information should be tailored to different grower's literacy and expertise.

If the above information about the user was available to a search system such as AgAsk, then the retrieval model could take this into account when ranking passages. In Telegram, this information could be recorded as part of a user's profile. How to use this information in one of the retrieval models (e.g. TILDEv2) is an open and interesting area of future work.

7.3 Deploying AgAsk in different regions

AgAsk was developed with users and data taken primarily from an Australian context. While many farming practices are universal, there are region specific characteristics. In particular, the industry reports indexed by AgAsk pertain to agriculture in an Australian environment. They are also all in English.

Our review of related work revealed India as being a region where technology solutions have been developed. How might one adapt AgAsk for deployment in, for example, India? We note that there is nothing region specific in terms of the underlying technology for AgAsk: the retrieval model, the Macaw chatbot framework and the Telegram client are all region agnostic. The language and region are determined by the collection of documents indexed in AgAsk. Thus, if a suitable collection of documents, containing information that users are interested in, can be compiled, then AgAsk can be deployed to serve this other region.

8 Conclusion

This paper presents AgAsk—a search system designed to answer farmers questions where information is extracted from scientific documents. While scientific documents on agriculture contain a plethora of useful information, they are not accessible or easily searchable by farmers with specific information needs. AgAsk attempts to overcome this by building a search system specifically for this problem.

Understanding the information needs of farmers is critical in designing a good search system to support them. We conduct a thorough analysis of information needs through a survey of users who were given real search scenarios to perform. This reveals the type of information they look for (e.g. crop protection, product recommendations) as well what source of information they use (books, Google, product sheets), what form they would like their answers as (e.g. a short answer with link to longer document). Learnings from this project informed the requirements for a search system and the basis of forming a test collection to evaluate such a system.

We form a test collection comprising 210 real questions, a collection of 86,846 scientific documents (split into 9,441,693 passages). Two agricultural experts did manual relevance assessment indicating which passages were relevant to each question. This provides ground truth for both training machine learning retrieval models and for empirical evaluation. The collection contains different query types (natural language vs. keyword), as well as human generated answers, thus providing a resource for further research on query variations and automated answer generation. The test collection is made public to foster further research into search in the agricultural domain.

Using the test collection we train and evaluate a number of passage retrieval models, including two state-of-the-art neural rankers—TILDEv2 and monoBERT. An empirical evaluation of all methods shows that neural rankers can be highly effective at finding relevant passages to a farmer's question.

How to deploy the above models in a usable system is often non-trivial. We describe a deployment architecture that makes use of the Telegram messaging platform for the front-end client and Macaw conversational search platform for the back-end server. This provides a flexible and scalable architecture. An analysis of the efficiency–effectiveness tradeoff of different retrieval models highlights how neural rankers such as monoBERT are not practical for deployment in live systems, and thus, alternative, non-GPU models such as TILDEv2 are preferred.

Finally, we highlight how the agricultural domain offers an interesting test bed for further research, with a key focus on better personalisation/contextualisation (e.g. location or weather aware rankers). It is our aim to both foster more

research in this area and to translate research into real-world systems deployed in the field.

Funding Open access funding provided by CSIRO Library Services. Funding for this project was provided by the Grain Research Development Corporation under project# UOQ2003-009RTX.

Data Availability Statement Data, code and the test collection are available at: https://github.com/ielab/agvaluate.

Declarations

Conflicts of interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval Ethics approval related to the survey we conducted was granted by The University of Queensland under application #2020000826.

References

1. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: ICLR (2017)
2. Bacco, M., Barsocchi, P., Ferro, E., Gotta, A., Ruggeri, M.: The digitisation of agriculture: a survey of research activities on smart farming. Array **3**, 100009 (2019)
3. Bailey, P., Moffat, A., Scholer, F., Thomas, P.: Uqv100: a test collection with query variability. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 725–728 (2016)
4. Bast, H., Korzen, C.: A benchmark and evaluation for text extraction from pdf. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 1–10. IEEE (2017)
5. Bendersky, M., Croft, W.B.: Discovering key concepts in verbose queries. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 491–498 (2008)
6. Chauhan, N.M., et al.: Information hungers of the rice growers. Agric. Update **7**(1/2), 72–75 (2012)
7. Chen, A., Liu, C.: Intelligent commerce facilitates education technology: the platform and Chatbot for the Taiwan agriculture service. Int. J. e-Educ. e-Bus., e-Manag. e-Learn. **11**, 1–10 (2021)
8. Craswell, N., Mitra, B., Yilmaz, E., Campos, D.: Overview of the trec 2020 deep learning track. arXiv preprint arXiv:2102.07662 (2021)
9. Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Lin, J.: Overview of the TREC 2021 deep learning track. In: Text Retrieval Conference (TREC). TREC, May (2022)
10. Hsu, C.-C., Lind, E., Soldaini, L., Moschitti, A.: Answer generation for retrieval-based question answering systems. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 4276–4282 (2021)
11. Jain, M., Kumar, P., Bhansali, I., Liao, Q.V., Truong, K., Patel, S.: Farmchat: a conversational agent to answer farmer queries. ACM Interact. Mob. Wearable Ubiquitous Technol. **2**(4), 1–22 (2018)
12. Jain, M., Kumar, P., Bhansali, I., Liao, Q.V., Truong, K., Patel, S.: Farmchat: a conversational agent to answer farmer queries. Proc. ACM Inter. Mobile Wearable Ubiquitous Technol. **2**(4), 1–22 (2018)
13. Jain, N., Jain, P., Kayal, P., Sahit, J., Pachpande, S., Choudhari, J., et al.: Agribot: agriculture-specific question answer system. IndiaRxiv (2019)
14. Kaszkiel, M., Zobel, J.: Passage retrieval revisited. In: ACM SIGIR Forum. vol. 31, pp. 178–185. ACM New York, NY, USA (1997)
15. Lipani, A., Losada, D.E., Zuccon, G., Lupu, M.: Fixed-cost pooling strategies. IEEE Trans. Knowl. Data Eng. **33**(4), 1503–1522 (2019)
16. Lipani, A., Palotti, J., Lupu, M., Piroi, F., Zuccon, G., Hanbury, A.: Fixed-cost pooling strategies based on IR evaluation measures. In: Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8–13, 2017, Proceedings 39, pp. 357–368. Springer, Berlin (2017)
17. Liu, X., Croft, W.B.: Passage retrieval based on language models. In: Proceedings of the eleventh international conference on Information and knowledge management, pp. 375–382 (2002)
18. Lokers, R., Knapen, R., Janssen, S., van Randen, Y., Jansen, J.: Analysis of big data technologies for use in agro-environmental science. Environ. Model. Softw **84**, 494–504 (2016)
19. Mills, J., Reed, M., Skaalsveen, K., Ingram, J.: The use of twitter for knowledge exchange on sustainable soil management. Soil Use Manag. **35**(1), 195–203 (2019)
20. Moffat, A., Scholer, F., Thomas, P., Bailey, P.: Pooled evaluation over query variations: Users are as diverse as systems. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1759–1762 (2015)
21. Momaya, M., Khanna, A., Sadavarte, J., Sankhe, M.: Krushi—the farmer chatbot. In: 2021 International Conference on Communication information and Computing Technology (ICCICT), pp. 1–6. IEEE (2021)
22. Nogueira, R., Yang, W., Cho, K., Lin, J.: Multi-stage document ranking with BERT. arXiv preprint arXiv:1910.14424 (2019)
23. Ogilvie, P., Callan, J.: Combining document representations for known-item search. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 143–150 (2003)
24. Opoku-Agyemang, K., Shah, B., Parikh, T.S.: Scaling up peer education with farmers in India. In: Information and Communication Technologies and Development, ICTD'17, pp. 15:1–15:10. ACM (2017)
25. Russell-Rose, T., Chamberlain, J., Azzopardi, L.: Information retrieval in the workplace: a comparison of professional search practices. Inf. Process. Manag. **54**(6), 1042–1057 (2018)
26. Salampasis, M., Fuhr, N., Hanbury, A., Lupu, M., Larsen, B., Strindberg, H.: Integrating IR technologies for professional search. In: European Conference on Information Retrieval, pp. 882–885. Springer, Berlin (2013)
27. Sanderson, M., et al.: Test collection based evaluation of information retrieval systems. Found. Trends Inf. Retr. **2**(2), 247–375 (2010)
28. Smith, M.J.: Getting value from artificial intelligence in agriculture. Anim Prod. Sci. 60(1) (2018). https://doi.org/10.1071/AN18522
29. Tait, J.I.: An introduction to professional search. In: Professional Search in the Modern World, pp. 1–5. Springer, Berlin (2014)
30. Teevan, J., Collins-Thompson, K., White, R.W., Dumais, S.T., Kim, Y.: Slow search: information retrieval without time constraints. In:

Proceedings of the Symposium on Human–Computer Interaction and Information Retrieval, pp. 1–10 (2013)
31. Tende, I.G., Aburada, K., Yamaba, H., Katayama, T., Okazaki, N.: Proposal for a crop protection information system for rural farmers in Tanzania. Agronomy **11**(12), 2411 (2021)
32. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. arXiv preprint arXiv:2104.08663 (2021)
33. Van Dalsem, S.: An iphone in a haystack: the uses and gratifications behind farmers using twitter. Master's thesis, University of Nebraska (2011)
34. Verberne, S., He, J., Kruschwitz, U., Wiggers, G., Larsen, B., Russell-Rose, T., de Vries, A.P.: First international workshop on professional search. In: ACM SIGIR Forum. vol. 52, pp. 153–162. ACM New York, NY, USA (2019)
35. Virgona, J., Daniel, G., et al.: Evidence-based agriculture—can we get there? Agric. Sci. **23**(1), 19 (2011)
36. Voorhees, E.M., Harman, D.K., et al.: TREC: Experiment and evaluation in information retrieval, vol. 63. Citeseer (2005)
37. Zamani, H., Craswell, N.: Macaw: an extensible conversational information seeking platform. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2193–2196 (2020)
38. Zhuang, S., Zuccon, G.: Fast passage re-ranking with contextualized exact term matching and efficient passage expansion. CoRR arXiv:2108.08513 (2021)