Conversational agent for aerospace question answering A position paper

ALEXANDRE ARNOLD, GÉRARD DUPONT, CATHERINE KOBUS, FRANÇOIS LANCELOT, and POOJA NARAYAN*, AIRBUS Artificial Intelligence Research

This paper introduces a research problem currently faced in the use of conversational agents within the aerospace industry. The aerospace domain is characterized by products and systems that are built over decades of engineering to reach high levels of performance within complex environments. This has resulted in technical documentation piling up over time.

The natural next step is to optimize the access to such document repositories to ensure completeness and relevance of information requested by operators. Recent developments in natural language processing, text understanding and information retrieval have provided major breakthroughs in the form of conversational interfaces for contextual search and cognitive assistants. However, the context of aerospace product documentation raises a number of specific challenges that are currently not fully addressed by advances in conversational agents intended for the general public.

This position paper describes some of these particularities. Even if these are not unique for the domain nor new, they serve to assess the recent progress in evaluation protocols and initiatives for conversational agents. Finally, we propose an evaluation protocol that may address these challenges.

CCS Concepts: • **Information systems** → *Enterprise search.*

Additional Key Words and Phrases: computation & language, information retrieval, question answering, conversational agent

ACM Reference Format:
Alexandre ARNOLD, Gérard DUPONT, Catherine KOBUS, François LANCELOT, and Pooja NARAYAN. 2019. Conversational agent for aerospace question answering A position paper . 1, 1 (July 2019), 6 pages. https://doi.org/NA

1 INTRODUCTION & CONTEXT

Large aerospace corporations rely on massive archives of documents, manuals and procedures that are often in paper format and are difficult to exploit. In most cases, these documents adhere strictly to the framework of aerospace regulations and certifications. The need for completeness and precision of the information sometimes enforces a structure in the document that makes it challenging for a digital assistant to fully exploit them. Each document talks about several related parts and is perfected over several evolution of aerospace systems creating various versions which creates complexity in terms of information flow and content types (see figure 1). A user looking for specific information related to a given scenario in this large corpus is often seen spending a significant amount of time navigating through the documents at hand. Even an experienced user who is familiar with the structure and template of the documents is faced with difficulties when information related to a specific product model or scenario is desired.

Search technologies are a way to address the structural complexity; however, they come with their own limitations. Most of the time, it is the user's responsibility to define their search needs through specific query syntax and refine the query until the right information is uncovered. For simple queries that have a ready-made answer in the document, this is not always a difficult problem. However, for the understanding of complex procedures or for troubleshooting system errors, it can lead to multiple queries thus a cumbersome search experience for the user.

The recent advances in natural language understanding and interactive search system have attempted to reduce cognitive overhead. The use of natural language conversation with the system can alleviate the users' need to understand the system's query syntax or document structure. Coupled with high-performing speech-to-text systems, it can even reduce the dependency to physical inputs to free users hand, allowing better multitasking. In this direction, conversational search agents appears to be a promising approach. For a more complete review on data driven dialog systems, please refer to [24] or [19] for a promising framework for conversational search.

Conversational search systems offer an interesting alternative to traditional document retrieval or navigation systems. However, the context of aerospace product documentation poses a number of challenges that are currently not fully addressed by classical conversational agents which target daily dialogues and conversations. Two antinomic objectives seems new: the need to collect as much context as possible to improve the accuracy of answers and the reduction of dialogues turns to obtain the answer to decrease cognitive workload of skilled operators. Moreover, in critical situations, the agent needs to maintain natural dialog with stressed humans while keeping the ability to identify with high confidence cases that are not understood (and revert to a secure alternative). For these reasons, the evaluation of conversational agent becomes key to enable their use in the aerospace domain.

In this paper, the authors explore existing evaluation frameworks and protocols for conversational agents in order to analyze their capabilities and limitations in the said aerospace context.

2 ASPECTS OF EVALUATION PROTOCOL

2.1 Specificity of aerospace conversational search

The case for a specific view on **Information Retrieval (IR)** and **Question Answering (QA)** for the aerospace domain offers a set a specific challenges for the conversational agents.

The source of knowledge - the document corpora - is in essence highly structured. However, this structure is mainly relevant to comply with regulations and processes and serve little purpose to improve retrieval of information associated to real world problems.

*Authors in alphabetic order.

Authors' address: Alexandre ARNOLD; Gérard DUPONT; Catherine KOBUS; François LANCELOT; Pooja NARAYAN, alexandre.arnold,gerard.dupont,catherine.kobus,francois.lancelot,pooja.narayan@airbus.com, AIRBUS Artificial Intelligence Research, 2 rond point Emile Dewoitine, PO Box D42-03, Blagnac cedex, France, 31703.

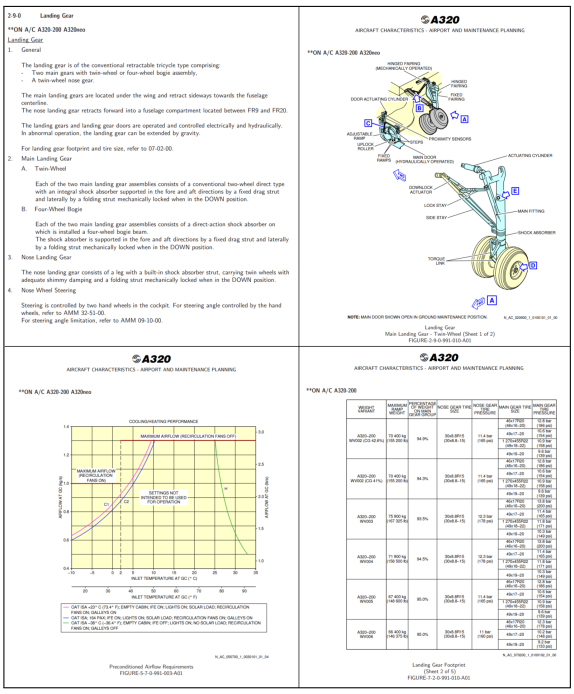

Fig. 1. Extracts from AIRBUS public documentation on **Airport and Maintenance Planning**. Full documents available at: http://www.airbus.com/aircraft/support-services/airport-operations-and-technical-data/aircraft-characteristics.html

In most cases, the content is not linear and contains many internal and external references to ensure completeness of the information provided. There are many technical diagrams and tables to ensure completeness but they do not facilitate fast interpretation by readers (see example in figure 1).

Also, in these documents, the relevant pieces of information relevant to specific tasks are often fragmented. The main reason is that the document has to cover numerous product variants and usage modes which superimpose the volume of information for each item (again see figure 1). Due to the large scope of products/variants covered, these documents are regularly updated with notes, addendums and sometimes amendments. Thus, the information is found in very long documents (hundreds to thousands of pages) through very diverse content (semi-structured text, images, diagrams, tables...). From an agent's point of view this can be considered as highly unstructured knowledge.

The aerospace IR/QA context is related to goal-oriented conversation where there is a problem to be solved related to a precise product and situation. However, the formulation of efficient analytical questions which summarize the issue at hand is often difficult. Thus, the problems are often formalized as a sequence of linked questions necessary to define the relevant context. In the end, finding a complete answer may need to link several pieces of information located in different parts of the document.

Finally, even if the end users are often qualified and even certified, comprehending the changes in the documents to quickly find answers is a skill that needs time and practice. Moreover, they are often operating in a time-critical environment to solve a problem impacting the efficiency of an aircraft flight or of a satellite constellation mission. The coherence of the conversation with the agent as well as the relative natural path taken in the discourses are important aspects to be assessed to ensure exactness of understanding.

2.2 Existing evaluation approaches

Most of the traditional "business metrics" for conversational agents are linked to web-based/B2C applications. The specificity of the aerospace context excludes the applicability of most of these classical approaches. Classical approaches mostly rely on retention and engagement metrics, related to the business goals and/or commercial funnel. Research in conversational agents has emerged from multiple academic domains in parallel and naturally each of the domains have adopted their own evaluation approaches. Among multiple proposals made, four major perspectives stand out in the literature.

Historically, the **Computer-science and AI perspective** came up with the now famous Turing-test (see [23] for a review). However, it does not involve a fine-grained measure of agent's performance and often completely ignores possible goals of the conversation. Even if more recent initiatives relate to the original test (see for instance [13]), it is mostly to overcome these limits.

Information Retrieval (IR) approaches are applicable to question/answering conversation. These focus on utility, relevance and timeliness aspects of the answers given to the user based on a problem statement. Precision, recall and related variants based on *hit@k* measures are all valid in this context. It does not encompass the conversation aspects but offer a clearer evaluation of strict utility.

Linguistic perspective (NLP) puts emphasis on conversation and cooperation with human discourse. In these evaluation protocols, the quality and coherence of the discourse are often weighted against a topical categorization of the conversations.

Finally, the **User experience (UX)** methodology is based on the human factors perspective and usability from the user point-of-view. This is the approach that better takes into account the end user perspective due to their direct involvement. However, it requires to setup interactive evaluations that can be time consuming and complex in terms of logistics, unless one already has an online system.

This categorization may not be the only possible one but seems relevant as a good baseline to define an evaluation protocol. Selection of one of the approaches depends on the conversational agent's final application and objectives. Hybridization of protocols is possible but alignment of evaluation metrics with overall business goals is necessary to get any proof-of-value.

3 EXISTING FRAMEWORKS

3.1 Conversational evaluation framework

In the following section, we present recent efforts in QA, text understanding and conversational agent evaluation frameworks. It is

a high-level overview, with some specific examples relevant to the special case of aerospace QA agent.

3.1.1 Overview.

- **Alexa prize** is a competition between academic laboratories organized by Amazon. Two events have been organized in 2017 and 2018 (see [21]) and another event has been launched for 2019. It encompasses the whole design, implementation and evaluation of social bots (meant to have open entertaining conversations) over a year of work with selected universities as participants.
  The task itself consists of voice-to-voice interactions and the competition offers a mix of evaluation and data collection through randomized human-to-bot conversations. An existing dataset of over 1 million conversations is distributed to the participants who are expected to develop their bots with the Alexa SDK. This framework presents a set of metrics that are generic enough to be applicable in many cases. One should note that these metrics were developed for the evaluation of social bots and often relate to the user experience methodology. These have been positively correlated to human judgments. They are also considered as an unification scheme that allows to rank the bots in the competition.
- **CommAI-env** (Environment for Communication-based AI) is an initiative from Facebook for training and evaluating AI systems (agents). A particular note is that CommAI does not make any assumption on language or the input modalities to the agent. The data exchange occurs as bit-level stream - even if the online running mode assumes the bits are characters in standard encoding for display purposes. The agents are trained by exposing them to various tasks and evaluated by considering the average reward accumulated by the agent just like in a Reinforcement Learning framework.
  This initiative has been criticized[1] due to its emphasis on computer science centric approach and bit-level communication (ie. turing machine approach) and thus trying to overcome NLU difficulties by working around it.
- **ParlAI** Facebook proposed ParlAI[15] an open-source framework that can be used for deployment and evaluation of conversational agents. It currently contains a lot of academic tasks, covering a broad scope of use-cases[1]: QA, goal-oriented chat (eg. booking), chitchat, negotiation, visual QA, cloze deletion test.
  Few of these tasks are really relevant to the aerospace QA use-case. The framework itself is however a huge effort in terms of evaluation standardization. It allows testing multiple approaches on multiple tasks within a unique environment. Among the most promising tasks, one can select the closed QA scenarios. In these scenarios, the agent has to answer questions, picking the response from one or several documents that are "sure" to contain the answer. The level of difficulty varies based on the documents that are used.
- **TREC - Conversational Assistance Track (CAsT)** The Conversational Assistance Track is a forum for building and testing systems that engage in open-domain information-centric conversational dialogues. The main aim of TREC CAsT is to advance research on conversational search systems. Just started in 2018, its first iteration will run in 2019. It aims at providing a standard open-domain reusable benchmark for Conversational Information Seeking (CIS) as the need was highlighted in the recent SWIRL report [7]. It builds upon existing work from the Dialog State Tracking Challenge [26], the QuAC [3] benchmark or the Wizard of Wikipedia [5] benchmarks and tries to extend the scope to large document collections with strict definition of information needs. Evaluations rely for now on relevance of selected information candidate with regards to topics (NDCG@K@N). It does not yet address coherence and dialog effectiveness but plans to extend to such more conversational aspects in later years including live experiments.

3.1.2 Synthesis. This overview of existing evaluation frameworks proposed in the literature allows to make a few observations (based on results summarized in Table 1):

- Most of these frameworks take their roots in the IR domain and its long tradition of structured evaluation protocols. It allows to ensure a fair comparison of approaches on stable document collection through optimized metrics.
- The UX approach is the less represented in the academic literature. This is natural due to the inherent complexity of interactive experiment with human subjects and the difficulty to compare results across such experiences.
- The NLP metrics are often cited as inspirations for half of these evaluation frameworks, however their interpretations are often limited in terms of generalization. Their use for information seeking or question answering has indeed be criticized (see for instance review on BLEU metric in [22]) and thus their are not common as main measures of performance.
- The AI approach seems to be used in very specific contexts. It is clearly suitable to the particular angle taken by CommAI-env. For Alexa-prize, the Turing test is expressively mentioned to not be a focus, but the long term goal of having a 20 min long coherent conversation indirectly follows the inspiration of a machine being indistinguishable from a human.

Table 1. Categorization of the evaluation framework against predefined types (bigger dot reflects stronger link).

Evaluation approach / type	AI	IR	NLP	UX
Alexa prize	·	·		●
CommAI-env	●		·	
ParlAI		●	·	
TREC - CAsT		●		·

Overall, the selection of the evaluation methodology for the aerospace use-case at hand seems to naturally align with the IR approach to start with. It ensures strict and reproducible performance evaluations that allow building systems incrementally while ensuring a stable baseline for performance comparison. The intrinsic dynamics of the conversational aspects are however not fully captured by such approach. The UX approach should thus offer a

[1] complete list http://parl.ai/static/docs/tasks.html

complementary view on system performances in user-centric experiments.

3.2 Focus on some tasks/datasets

In the following, we present a selection of existing evaluation tasks which are possible surrogates or precursors for aerospace QA.

- **SQuAD** Stanford Question Answering Dataset [20] questions collected from crowd-workers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage. It contains 100k+ question-answer pairs on 500+ articles. There is however no notion of dialog nor construction of response in natural language and the questions are mostly fact-based.
- **bAbI tasks** [25] provide a set of artificial tasks, in a similar way to how software testing is built in computer science. Initiated by Facebook, the aim is that each task tests a unique aspect of text and reasoning, and hence tests different capabilities of learning models. Each QA instance consists of a set of sentences stating facts, a single factual question, its answer and supporting fact(s). Conversation aspects are not the focus here in favor of identifying key QA and reasoning capabilities. Supporting text and question are by definition artificial which allows control of each incremental experiments.
- **CLEVR** [9] is a diagnostic dataset for Compositional Language and Elementary Visual Reasoning (CLEVR). Questions in CLEVR aim at testing various aspects of visual reasoning including attribute identification, counting, comparison, spatial relationships, and logical operations. Each question in CLEVR is represented both in natural language and as a functional program. The dataset consists of generated images with roughly 10 associated questions and answers. Each scene also has its own graph annotation giving ground truth of attributes, objects and relations as well as a functional programmatic representation of the reasoning skill required to answer questions. The code for generating the CLEVR dataset is available on GitHub. If the artificially generated dataset can be very far from real case, this dataset offers a unique way to test image understanding in a very controlled environment.
- **MS MARCO** or Microsoft MAchine Reading COmprehension Dataset [16], is a large scale dataset for reading comprehension and question answering. It comprises a large set of more than 1M real user queries from Bing search engine, more than 100k natural language answers or no answer subset over 10 text passages per query. It addresses the issue of text understanding and natural language response generation from an open domain set (the Internet). It allows to test a model at scale, but this openness of the text collection and the constraints applied on the responses set make the understanding of performance quite limiting.
- **WikiQA** is a set of question and sentence pairs, collected and annotated for research on open-domain question answering. Each question is linked to a Wikipedia page that potentially has the answer. It includes 3,047 questions and approximately 30k sentences in the dataset, where 1,473 sentences were labeled as answer sentences to their corresponding questions.
- **Movie Dialog dataset** goal and non-goal oriented dialog centered around the topic of movies [6]. It combines 3 types of tasks: question answering, recommendation and discussion. It also has a task that is combining the 3 first. This dataset presents very interesting characteristics: a closed repository of knowledge (database of movie facts) to answer factoid questions, a set of recommendations per movie facet, a set of real world dialogues and finally a compilation of real world discussions on movies (from Reddit). If the proposed tasks are somehow limited compared to a generic aerospace use-case the completeness of the dataset makes it quite relevant.
- **Ubuntu dialog** corpus [14] offers 1M dialog with roughly 8 turns each (on average, with a minimum of 3 turns and a few cases with dozens of turns) between 2 users. It comprises 7M utterances using 100M words of vocabulary. Highly technical and goal (problem-solving) oriented, these dialogues offer a proper evaluation environment for "tech support" bots. The text content is real but the forum structure does drive the sequences of utterances away from a natural conversation and makes it difficult to exploit.
- The **Textbook Question Answering** (TQA) dataset is drawn from middle school science curricula as described in [10]. It consists of 1, 076 lessons from Life Science, Earth Science and Physical Science textbooks. Each lesson has a set of multiple choice questions that address concepts taught in that lesson. TQA has a total of 26k questions including 12k that have an accompanying diagram.
- **SimpleQuestions**[2] is a dataset for simple QA; it consists in a total of more than 108k questions, written in natural language by human annotators. Each question is paired with a corresponding fact, formatted as a triplet (subject, relationship, object) that provides not only an answer but also a complete explanation.
- **Children's Book Test (CBT)**[8] is a dataset (training size of more than 669k questions) built from books that are freely available. In this dataset, 'questions' from chapters are built by enumerating 21 consecutive sentences. The first 20 sentences form the context while a word is removed from the 21st sentence (becoming the query). The missing word must be identified among a selection of 10 candidate words appearing in the context or in the query.
- **HotpotQA** is composed of 113k QA pairs based on Wikipedia; questions are designed to require multi-hop reasoning, i.e. searching for the answer in different sources. The queries are also highly diverse and not limited to any pre-existing knowledge base. This dataset also relies on strong supervision for supporting facts, which enables more explainable question answering systems. The multi-hop reasoning aspect of the task is key for the aerospace documentation where in one documentation, you can have links/references to other ones and you might need to pick information in those to build the answer.
- **Google's Natural Questions** is a QA dataset, whose questions consist of real users' queries issued to the Google search engine [11]. An annotator is presented with the top 5 search

results; for each (question, Wikipedia page) pair, the annotator returns a (long answer, short answer) pair. The training set contains 307k examples with single annotation while development and test set contain about 7,8k examples with 5-way annotations.

- **QuAC** [3] provides 14k dialogues that encompass questions over Wikipedia. The definition of questions and answer is similar to the SQUAD dataset, however most of the sequences are to be taken in context with multiple references and coreferences of entities from one question to another. The conversation is however limited to question/answer pairs without clear flexibility in the dialog flow. If the context aspects makes this dataset challenging, the dialogues where constrained by a set of only 3 dialog-acts to answer any question: (1) continuation (follow up, maybe follow up, or do not follow up), (2) affirmation (yes, no, or neither) and (3) answer-ability (answerable or no answer).
- **MSDialog** [17][27][18] dataset is a labeled dialog dataset of question answering (QA) interactions between information seekers and answer providers from an online forum on Microsoft products (Microsoft Community). The dataset contains more than 2,000 multi-turn information-seeking conversations with 10,000 utterances that are annotated with user intent on the utterance level. Annotations were done using crowdsourcing with Amazon Mechanical Turk. MSDialog has several versions, including the complete set (MSDialog-Complete) and a labeled subset (MSDialog-Intent).

These initiatives range from classic QA to more abstract text understanding and even pure conversation tasks. A hybridization of these appears necessary to encompass an end-to-end evaluation of aerospace QA (see section subsection 4.2).

4 AEROSPACE CONVERSATION SEARCH EVALUATION

The broad scope of conversational search agent evaluation initiative is encouraging. Given the relative recency of such initiatives, their adaptation to more application domain is still limited. The aerospace QA domain particularities arise out of the following facts:

- Aerospace use-case is a closed domain where the source of knowledge is a fixed set of documents;
- Knowledge is organized in a semi-structured fashion in long documents with hundreds of pages including substantial number of complex graphics and nested tables ;
- The task of question/answering has a large scope due to variations of possible questions, multi-hop nested tasks ;
- The user population has a good knowledge of the domain but is under pressure for finding timely and secure response to the problem at hand.

Even if the scope of the domain is closed, the volume of data manipulated and the large space of possible questions will not permit to have manual generation of questions/answers with decent coverage. The importance of question context which is related to product version and variants increases the need for a real conversation between the agent and the user in order to alleviate any uncertainty regarding the understanding of scenario being dealt with.

Finally, the particular case of answering questions about such complex systems leads to an imbalanced cost of errors. It is even increased by a tolerance to errors which will vary depending on the required sensitivity and specificity for a given task. A noteworthy example of Aircraft safety related question/answering could be quoted here where no error could be tolerated.

4.1 Metrics

Since the context of the conversation does not favor a particular dialog approach, the metrics used should be independent from the agent approach. Thus, as proposed in [12], model-independent metrics should be preferred and in particular those that do not suppose any supervised signal.

The need for precision tends to favor traditional information retrieval metrics (precision, recall, f-measure[4]). Regularly used in most QA and search tasks, these measures allow an absolute measurement of performance of the agent. They need an unambiguous supervision signal given as a set of search tasks and exact responses which could be tedious to build.

Moreover, the objectivity and correlation to human perception of this precision is however to be confirmed. To overcome this limit, the Alexa prize proposes some specific metrics:

- **engagement** through number of dialog turns and conversation duration
- **coherence** measured through Response Error Rate: $RER = \frac{number\ of\ non\ coherent\ responses}{number\ of\ utterances}$
- **conversational depth** as the average number of consecutive turns on the same topical domain

These have been shown to correlate positively with human perception (see [21]) of Conversational User eXperience (CUX - collected through user feedback in large scale interactive evaluation). To be able to compute some of these metrics, an additional set of measurements related to topical variation of the conversation is proposed:

- topic classification (among 26 predefined topics for Alexa - to be adapted to aerospace domain)
- topical diversity/conversational breadth with topical vocabulary size and distribution of each topic
- domain coverage captured by measuring the entropy across distribution of number of conversations across domains

If engagement - as defined by the number of dialog turns - does not seem well suited for aerospace conversational search use-case, some proxies could be found by analyzing users responses depth and timings. As for task definition, there is no clear metric that can encompass the multiple aspects of aerospace QA. A combination of the ones listed here should allow 1) to offer a multi-faceted performance measure, 2) align with existing initiatives to compare different baselines and 3) offer a more human contrast to the purely IR performance metrics. Simple task based metrics (time to find an answer or solve a particular issue) will also be added to complete the view on the system usability.

Finally, in some specific operators environment (ie subject to stress factors), additional human factors measures related to information presentation will be critical metrics of the agent. We will however keep these aspects for further analysis.

4.2 Towards an evaluation protocol

To allow a clear assessment of agent performance, the best approach would obviously be to redefine large scale evaluation tasks on the specific data related to the aerospace problem. However, the amount of effort necessary to achieve this is high. If done without clear links to existing approaches, it would not easily permit comparison of agents' performances in other contexts. The goal is thus to be able to correlate performance in specific aerospace tasks with the ones tested in the open context of existing initiatives and identify promising approaches to be evaluated on specific aerospace tasks.

The current foreseen approach is then straightforward:

(1) assessing performances on existing evaluation frameworks ;
(2) defining and run context specific evaluation tasks on aerospace QA datasets (to be created) ;
(3) assess the correlation of performances between these ;
(4) run interactive human evaluation to confirm perception of system performance by the future operators on aerospace specific tasks.

A combination of information retrieval metrics with the dialog oriented ones listed above seems necessary to measure the human perception of the agent performance. As far as the aerospace QA constraints have been defined, the closest evaluation task is the Movie Dialog dataset which combines: goal and non-goal oriented aspects within a closed domain of knowledge. It should serve as a valuable start and as a possible methodology in building new dedicated evaluation tasks. In terms of evaluation framework, ParlAI appears to allow a solid structure for deployment and evaluation.

This approach should offer a proper evaluation context for aerospace conversational agent and a point of reference serving as baseline to improve on. In the long run, the necessity to complement this evaluation proposal with more interactive tasks - and thus getting toward more realistic search tasks - will be necessary. Long-standing literature exist on the subject, especially in the context of the TREC conferences but this is currently out of the scope of this paper.

5 SUMMARY AND FUTURE WORK

The present position paper introduced the difficulties faced in applying conversational search agent within the context of aerospace question answering. Complexity of the documents, high variability of the possible questions and imbalance of response error costs are among the most differentiating aspects compared to the literature and existing benchmark proposals. These difficulties are not sufficiently covered by the current evaluation frameworks to ensure that the performances of an agent measured in such context can be transferred in the aerospace context.

This is obviously a very preliminary and work in progress related to a ongoing research activity within AIRBUS. Our longer-term goal is to be able to develop more advanced evaluation protocols that help address these problems and facilitate research towards reliable conversational agent that can support the aerospace industry.

REFERENCES

[1] Gemma Boleda. 2016. Remarks on the CommAI-env. (2016). https://mainatnips.github.io/ Machine Intelligence Workshop @ NIPS 2016.

[2] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. [n. d.]. Large-scale Simple Question Answering with Memory Networks. ([n. d.]). arXiv:cs.LG, cs.CL/http://arxiv.org/abs/1506.02075v1

[3] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC : Question Answering in Context. arXiv:cs.CL/1808.07036

[4] W Bruce Croft, Donald Metzler, and Trevor Strohman. 2010. *Search engines: Information retrieval in practice*. Vol. 283. Addison-Wesley Reading.

[5] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations*.

[6] Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2015. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931* (2015).

[7] Allan et al. 2018. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum* 52, 1 (Aug. 2018), 34–90. http://doi.acm.org/10.1145/3274784.3274788

[8] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. [n. d.]. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. ([n. d.]). arXiv:cs.CL/http://arxiv.org/abs/1511.02301v4

[9] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *CVPR*.

[10] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are You Smarter Than A Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension. (2017).

[11] Tom Kwiatkowski et al. 2019. Natural questions: a benchmark for question answering research. (2019).

[12] Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. [n. d.]. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. ([n. d.]). arXiv:http://arxiv.org/abs/1603.08023v2

[13] Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. [n. d.]. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. ([n. d.]). arXiv:http://arxiv.org/abs/1708.07149v2

[14] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. [n. d.]. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. ([n. d.]). arXiv:http://arxiv.org/abs/1506.08909v3

[15] Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. [n. d.]. ParlAI: A Dialog Research Software Platform. ([n. d.]). arXiv:cs.CL/http://arxiv.org/abs/1705.06476v4

[16] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).

[17] C. Qu, L. Yang, W. B. Croft, J. Trippas, Y. Zhang, and M. Qiu. 2018. Analyzing and Characterizing User Intent in Information-seeking Conversations.. In *SIGIR '18*.

[18] C. Qu, L. Yang, W. B. Croft, Y. Zhang, J. Trippas, and M. Qiu. 2019. User Intent Prediction in Information-seeking Conversations. In *CHIIR '19*.

[19] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval - CHIIR '17*. ACM Press, Oslo, Norway.

[20] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. [n. d.]. SQuAD: 100,000+ Questions for Machine Comprehension of Text. ([n. d.]). arXiv:cs.CL/http://arxiv.org/abs/1606.05250v3

[21] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. [n. d.]. Conversational AI: The Science Behind the Alexa Prize. ([n. d.]). arXiv:http://arxiv.org/abs/1801.03604v1

[22] Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics* 44, 3 (2018), 393–401.

[23] Ayse Pinar Saygin, Ilyas Cicekli, and Varol Akman. 1999. Turing Test: 50 years later. *Minds and Machines* (1999), 2000.

[24] Iulian Serban, Ryan Joseph Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A Survey of Available Corpora for Building Data-Driven Dialogue Systems. *CoRR* abs/1512.05742 (2015).

[25] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. [n. d.]. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. ([n. d.]).

[26] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*.

[27] L. Yang, M. Qiu, C. Qu, J. Guo, Y. Zhang, W. B. Croft, J. Huang, and H. Chen. 2018. Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems. In *SIGIR '18*.