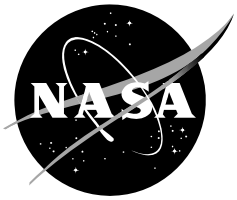
NASA/CR-2018-220097



NASA Pilot-Engaged Expert Response using IBM Watson Technology

Prototype Evaluation of Knowledge Retrieval System
Final Report

Graham Katz, Chengmin Ding, and Andrew Doyle
IBM Corporation, Herndon VA

October 2018

NASA STI Program ... in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counter-part of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- E-mail your question to help@sti.nasa.gov

- Phone the NASA STI Information Desk at 757-864-9658

- Write to:
  NASA STI Information Desk
  Mail Stop 148
  NASA Langley Research Center
  Hampton, VA 23681-2199

NASA/CR-2018-220097

NASA Pilot-Engaged Expert Response using IBM Watson Technology

Prototype Evaluation of Knowledge Retrieval System
Final Report

Graham Katz, Chengmin Ding, and Andrew Doyle
IBM Corporation, Herndon VA

National Aeronautics and
Space Administration

Langley Research Center
Hampton, VA 23681

October 2018

Acknowledgments

This work pursues NASA's desire to evaluate the feasibility of developing a Pilot Expert Advisor System for use in the flight deck concept which would monitor, advise, and assess in real-time and in partnership with the human to ensure safe and efficient operations and to overcome current-day automation failings.

The work was sponsored by the Airspace Operations and Safety Program, Autonomous System project, led by Ms. Sharon Graves, executed under the auspices of the NASA Langley Research Center (LaRC) Flight Deck Interface Technologies Team, led by Mr. Randall Bailey. The contributions of these individuals and organizations as well as those of Ms. Lisa Le Vie, Mr. Trey Arthur, Dr. Jon Holbrook, and Mr. Vincent Houston are gratefully appreciated.

Available from:

NASA STI Program / Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199
Fax: 757-864-6500

Table of Contents

List of Figures

List of Tables

Acronyms

CDA	Contextual Data Augmentation
CDT	Contextual Data Type
EFB	Electronic Flight Bag
IoT	Internet of Things
LaRC	Langley Research Center
NLP	Natural Language Processor
PEER	Pilot-Engaged Expert Response
QA	Question-Answer
QRH	Quick Reference Handbook
SME	Subject Matter Expert
STT	Speech-To-Text
TTS	Text-To-Speech
UI	User Interface
WDA	Watson Discovery Advisor

Executive Summary

NASA Langley Research Center and IBM have been investigating the use of IBM Watson technology in aerospace research and development. One application of Watson technology is the Pilot-Engaged Expert Response (PEER) use case. The PEER system is envisioned as an in-cockpit advisor that will act as a source of situationally-relevant information for pilots and other flight crew members to assist in decision making about real-time events and situations that arise in the course of aircraft operations. PEER will make available vast stores of knowledge and information quickly and directly, putting important informational resources where they are needed most.

IBM has worked with NASA to develop an architecture and articulate a roadmap for the development of the PEER system. That vision is built around Watson Discovery Advisor (WDA) software solution, derived from IBM's Jeopardy!-winning automatic question answering system. PEER makes use of WDA's sophisticated question-answering capabilities as its core, adding important User Interface components and other customizations for the cockpit environment, including communication with flight systems and other external data sources. The development plan for PEER includes four development stages, with the current project constituting the first phase.

In this project, a prototype instance of PEER was successfully adapted to the aviation domain, enabling users to ask questions about aviation topics and receive useful and accurate answers to these questions. Major tasks accomplished include the development of procedures for domain adaptation through automatic lexicon extraction from domain glossaries; generation of question-answer training data which was used to train the system; and assessment of the effectiveness of domain adaptation, which showed a dramatic improvement in the ability of the PEER system to answer domain-relevant questions. In addition, the vision for the PEER system was pushed forward by the articulation of a plan for the automatic enhancement of question-answering with contextual information.

This initial phase focused on two main goals: 1) the targeted domain adaptation of the underlying WDA system to the aviation domain; and, 2) the design of the software systems needed to leverage flight-contextual data.

Domain adaptation of the WDA system proceeds via three main activities: Domain data ingestion, lexical customization and model training. A textual corpus consisting of 1,147 individual documents with more than 7.5 million words of text was ingested into the system and this served as the basis of all further development. A domain lexicon of over 3,500 aviation-domain terms was semi-automatically generated from domain documents and used to train the system. In addition, a set of over 500 question-answer (QA) pairs relevant to the PEER use case was developed; these were used to train and assess the system. These important first steps established the basis for the PEER system.

In addition, steps were taken towards the integration of the PEER system into the cockpit environment with the development of a functional design for the Contextual Data Augmentation (CDA) subsystem. This subsystem brings to bear contextual data to improve system responses. It has three main sub-modules: the Contextual Data Collection module, the Contextual Data Selection module, and the Contextual QA Augmentation module. These modules form a processing pipeline that addresses the problems associated with automatically integrating information from external resources into the knowledge-retrieval mechanism.

Introduction

NASA Langley Research Center (LaRC) and IBM have been investigating the use of IBM Watson technology in aerospace research and development, particularly as this technology can help to leverage "big data" for mission applications. One promising application of Watson technology is its use as part of a pilot expert advisor system. The pilot expert advisor system is envisioned as an in-cockpit cognitive advisor for commercial transport aircraft flight crews. The system will be a source of situationally-relevant information for pilots (and other flight crew members) to assist in decision making about real-time events and situations that arise in the course of aircraft operations by making available vast stores of knowledge and information quickly and directly.

Modern commercial aircraft are heavily instrumented and extensively documented, as are aircraft technology, aircraft operations and operation history in general. In recent years much of the aircraft documentation has become available to the flight crew in electronic format in the form of the Electronic Flight Bags (EFB), electronic checklists, and Centralized Alerting and Warning Systems. These systems provide for the rapid availability of structured emergency (non-normal) response information, such as that provided by the pilot's Quick Reference Handbook (QRH). These responses, however, are scripted, based upon clearly identified fault conditions. They are typically limited to the most probable failure effects, and only identify actions responsive to these identified conditions that best minimize their effects. By the flight crew following the QRH, the aircraft should remain safe, but the process is not designed to assist the flight crew with troubleshooting, fixing the system, or conveying detailed systems knowledge or understanding.

This use of automation and prescribed response methodologies sets up the "automation paradox" (Harford, 2016). The paradox arises because automation accommodates incompetence in a user; it is easy to operate and can mask operator limitations. New operators (i.e., pilots) can quickly learn and be sufficiently proficient to operate an aircraft by use of automation. At the same time, if operators are competent or an expert, the use of automation erodes their manual operating skills. The use of automation does not allow the operator (i.e., pilots) to exercise their manual flying skills. The net result – the paradox – is that, since automation tends to fail in unusual situations or creates unusual situations, the inexperienced never acquire, and the experts lose the very skills required to 'save the day' when automation fails. The pilot expert advisor system described in this document – the Pilot-Engaged Expert Response (PEER) system – is intended to fill the knowledge and understanding gap as well as to support the human in developing and maintaining automation and aviation skills.

The PEER system is envisioned to be a cognitive system available to the flight crew at all times which is able to locate information relevant to flight crew queries posed in response to an ongoing flight event. By bringing to bear vast stores of aviation documentation, including technical documentation about the aircraft, historical records of aviation incidents, and general knowledge about aviation best practices, the PEER system will support the crew's situation awareness and understanding. IBM has worked with NASA to develop an architecture and articulate a roadmap for the development of the PEER system. That vision is built around the WDA software solution. WDA is derived from IBM's Jeopardy!-winning automatic question answering system Watson (Ferrucci, 2012). It can read and extract knowledge from a large collection of text documents and use that knowledge to respond to natural language questions posed to the system. PEER makes use of WDA's sophisticated question answering capabilities as its core, adding important User Interface components and other customizations for the cockpit environment.

The Watson knowledge retrieval and storage components of the PEER system are crucial to advising flight crews. There is, in addition, a range of additional functional components needed for the whole pilot advising system to be effective, including user interface components and data interpretation components. In this document, we describe the first phase of development for the PEER system, which was strictly

concerned with issues related to access to knowledge, and not with how this knowledge is presented users or otherwise put into effect.

This phase focuses on two main goals: 1) targeted adaptation of the underlying WDA system for the aviation domain; and, 2) design of the customizations needed to leverage contextual information to improve system responses.

Background

The Watson Discovery Advisor (WDA) is a cognitive computing system that orchestrates an ensemble of search technologies, natural language processing algorithms, and machine learning models, enabling users to find information in a large body of unstructured natural language text by posing simple, natural-language questions to the system (Ferrucci et al, 2010). WDA responds to natural language queries by finding documents and relevant pieces of documents in the collection, identifying and extracting likely answers to questions, and scoring and ranking those answers on the basis of their likelihood of being the correct answer.

The desktop user interface to the WDA system, illustrated in Figure 1, displays some of its features: The natural-language query, the top-ranked answers to the query ("Hypotheses"), and the "Key passages" supporting those hypotheses.

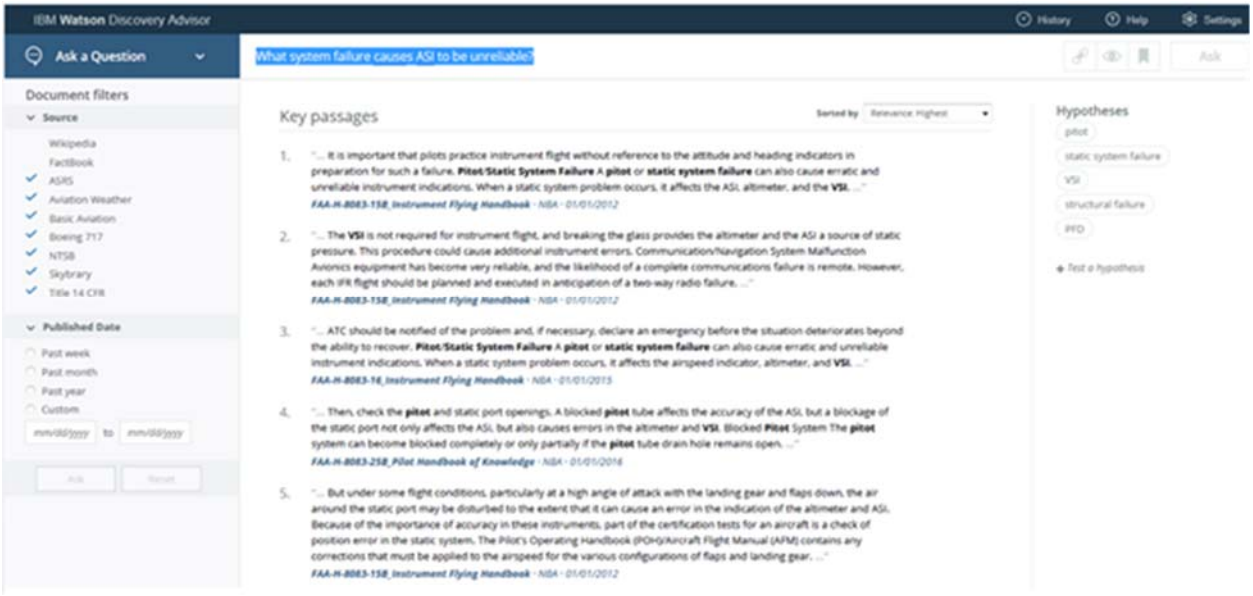


Figure 1: Watson Discovery Advisor User Interface

The WDA processing pipeline consists of five major components characterized below.

- **Question Analysis**: Applies Natural Language Processing (NLP) to input question to identify the semantic type of the answer sought and to generate text-retrieval queries based on the question
- **Corpus Search**: Uses the text retrieval queries to identify and rank passages from a document collection that are responsive to the query
- **Answer Generation**: Identifies words or phrases from passages that have the potential to be an answer to the target question, based on NLP analysis of the responsive passages and question
- **Answer Ranking**: Applies analytics to score answer hypothesis and rank them with trained ranking models based on the match between the responsive passages and the question
- **Display**: Displays top-ranked answer hypotheses and supporting passages for them

This question-answering capability is the core capability of the WDA system, and will form the core of the PEER system. To achieve the PEER vision, important user interface components will be built around the WDA core, enabling, for example, spoken language interaction and facilitating integration into the cockpit environment. Additionally, the functionality of WDA will be enhanced to address difficulties specific to the PEER Use Case, such as the interpretation of semi-structured text and causal-chain

reasoning. The PEER system will interact with aircraft systems and other informational sources to acquire information that will allow it to better address flight crew informational needs. The envisioned operational context for the PEER system is illustrated in Figure 2. When fully built out, PEER will provide a service for commercial aviation in general, both to flight crews and to other users.

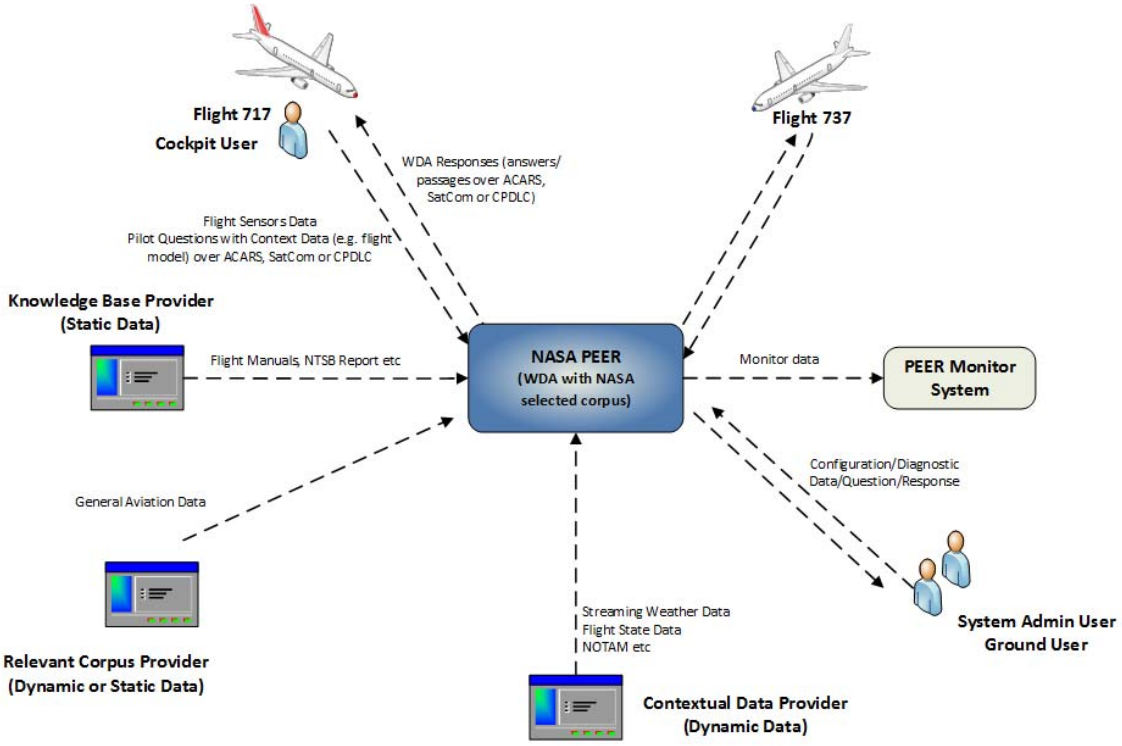


Figure 2: Pilot-Engaged Expert Response system context

IBM has worked with NASA LaRC to develop a phased approach to the realization of this vision for PEER. An initial Proof-of-Concept was built to address a specific scenario. This was an incident on May 12, 2005 in which a Boeing 717-200 operated by Midwest Airlines, Inc., as Flight 490, experienced a series of pitch oscillations related to unreliable airspeed indication while climbing to cruise altitude over Union Star, Missouri. The flight diverted to Kirksville Regional Airport (IRK), Kirksville, Missouri, after the flight crew declared an emergency (National Transportation Safety Board, 2018). Like all accidents and incidents, a variety of errors or circumstances had to all align for the event to occur. During the incident, a number of symptoms were evident, and in fact, identified by the flight crew, but they never fully identified their significance.

The initial task was to determine whether Watson could have provided the flight crew with information that would have assisted or prevented this loss-of-aircraft state awareness incident from occurring. A dataset of documents relating to this use case, consisting of general aviation documentation and documentation about the specific aircraft, was read into the WDA system. The prototype system was demonstrated to answer questions relevant to the target Use Case about the potential causes of the unreliable airspeed indications, identifying the pitot-static system as a likely source, with icing of the pitot port as a potential cause for the observed erratic instrument readings. This result was encouraging.

As part of this Proof-of-Concept, the vision for PEER was more fully articulated, an architecture for the PEER system was developed and a phased development plan proposed for the complete build out of the system. The PEER system was designed to have three major components – a User Interface, an Aircraft Interface and a (WDA-based) Discovery backend. This architecture is illustrated in Figure 3.

In all, the design encompasses 16 functional modules organized into six different architectural layers. This includes a WDA-based Deep Research Service and the Aircraft Adapter Service, augmented by other core services including Infrastructure Services, Information Integration Services, Base Supporting Services and Information Delivery Services.

The basic idea of the phased development plan was to build out from the Discovery core. The PEER implementation plan (shown in Figure 3 using color-coding as described in the following) involves additional articulation and development of each of the components in four planned phases:

Phase 1: Discovery, Query Building and Response Structuring (Yellow, Fig. 3)

Phase 1 will provide a limited initial end-to-end implementation of the PEER concept, illustrating that Watson technology can identify information appropriate to responding to flight situations for a particular aircraft type. PEER will draw on technical manuals and other resources, finding the applicable segments of text and delivering them to users on the basis of queries initiated by the flight crew, supplemented by context, or on a semi-automated basis, as triggered by particular automatic alerts. The current project constitutes the first part of Phase 1.

Phase 2: Enhanced Human Factor, Dynamic Context, and Measurements (Light Green, Fig. 3)

Phase 2 will enhance the user interface through human factors by accepting human-initiated queries in the form of voice and delivering information in the form of speech as well as text. Further articulation of the contextual augmentation of questions will be pursued through dynamic context information in the form of temporally-specific measurement data, such as the current weather or key current instrument readings, into the query and into the information discovery mechanisms.

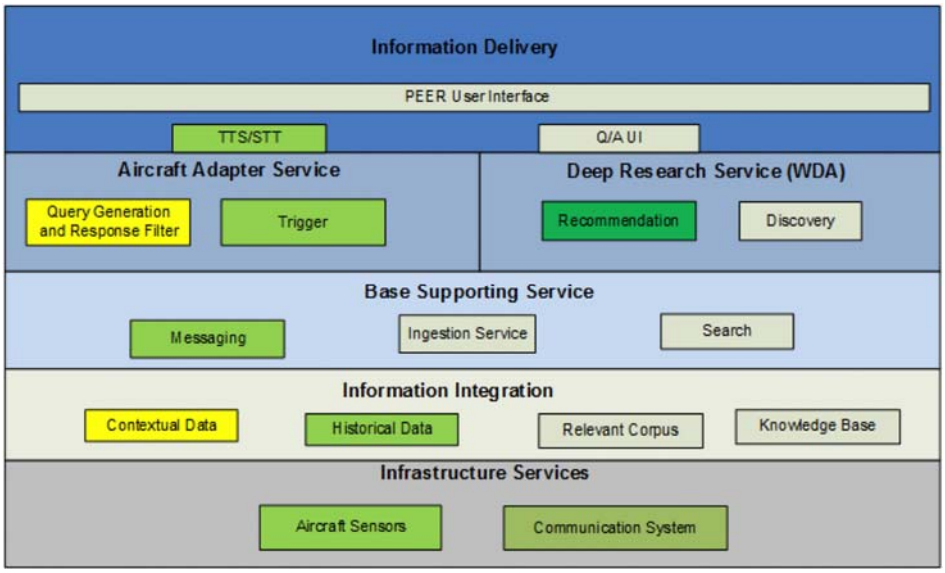


Figure 3: NASA PEER High-level Conceptual Architecture

Phase 3: Semantic Response Enhancements (Dark Green, in Fig. 3)

Phase 3 will expand upon the query response mechanism to support causal chain reasoning and incorporate explicit evaluative information into the discovery mechanism. PEER will break down questions into sequences of information gathering queries, whose results are combined to provide a response. In addition, techniques will be developed to ingest structured data consisting of tables and other explicit document-conveyed structured information.

Phase 4: Automatic Domain Adaptation Strategy

In this final phase, a strategy for extending the PEER system will be developed. This phase will include the development of a stepwise methodology for expanding the PEER's range of applicability outside the initial domains and implementing the initial steps of that domain adaptation strategy. Development of the strategy will involve identifying natural clusters of data and related use cases and organizing these into a logical sequence.

This overall development plan encompasses a wide range of enhancements and add-ons to the existing Watson technology, as well as PEER-specific customization.

Project Tasks and Goals

The current project is more narrowly focused and involved two major tasks: The first task was a more extensive adaptation of the WDA system to the aviation domain, including providing the NASA team with the ability to interact with the WDA User Interface and to pose questions, as well as an assessment of the effectiveness of the system. The second task was the development of design enhancements of the PEER system to make use of contextual information.

In summary, the central goals of this project were: 1) to adapt the Watson Discovery Advisor system to the aviation domain through lexical customization and model training and testing, and, 2) to design a system for incorporating contextual information into the WDA question-answering system pipeline.

Domain Adaptation

Domain adaptation of the WDA system proceeds via three main activities: 1) Domain data ingestion; 2) lexical customization; and, 3) model training.

Domain data ingestion involves identifying textual data sources and ingesting them into the WDA system. This was done as part of the initial Proof-of-Concept.

Lexical customization refers to the development of Watson domain dictionaries ("lexicons") to provide domain specific information to the Watson system. These lexicons provide Watson with information about a term's domain specific meaning, forming the basis for the system to identify answers to questions. The size of this domain vocabulary for the aviation domain is vast and the more complete the domain lexicon is, the more likely it is that WDA will be to answer domain questions directly. The goal of the lexical domain adaptation is to teach Watson enough about the language of aviation that it can address aviation-specific questions.

The final aspect of the domain adaptation is the training and testing of the WDA system using pairings of question with their correct answers (QA sets). This involves the development of the QA, training of the machine-learning ranking models, evaluating and testing the system, and analyzing sources of errors.

These domain adaptation activities are discussed in the *Domain Adaptation* section below section below.

Contextual Augmentation Design

In order for a system such as PEER to generate accurate and contextually-relevant responses to naturally formulated queries from the flight crew, the system requires a mechanism for augmenting queries to the system with contextual information. Contextual information plays a central role in identifying information relevant to flight crews. Contextual information can include information about the state of the aircraft and of aircraft systems, the current flight plan, the air traffic situation, what the weather is like and so on. For this project, a second goal was to develop a functional design for augmenting the question/answering

system with contextual information. This design is laid out completely in a separate document, but it is summarized in the *Contextual Data Augmentation Design* section below.

Domain Adaptation

Natural language processing systems such as WDA acquire knowledge from textual documents meant for human consumption by parsing and semantically analyzing the natural language sentences. As normally configured, WDA has NLP capabilities tuned to general interest questions and configured to process standard American English.

Domain Adaptation is the process of optimizing the WDA processing pipeline for a specific set of data and for the concerns of a specific community of users. In the case of PEER, this involved optimizing WDA for reading text about aviation and answering questions that the flight crew might pose to the system in the case of a flight incident. The first task of domain adaptation is to identify and read into the system a set of documents which is appropriate to the domain of application for the system. The second task is customization of the WDA system's processing pipeline.

Each of WDAs processing components (Question Analysis, Corpus Search, Answer Generation, Answer Ranking) makes use of elements that can be customized to the domain to which WDA is being applied. Besides textual corpus ingestion, the major tasks of domain adaptation are the customization of the NLP processing through the development of custom domain lexicons and the customization of the answer ranking models through training of the system with question/answer sets. Although both aspects of customization have wide-ranging effects, lexical customization tends to improve Corpus Search and Answer Generation while the model training improves Answer Ranking.

Corpus Textual Ingestion

For the PEER Use Case, a set of domain documents was identified by NASA as relevant. These documents came from seven different data sources and included both general aviation information and aircraft-specific information. The corpus followed the training principals for a commercial transport aircraft pilot: a) general aviation knowledge and experience sufficient to become licensed; and, b) detailed, type-specific knowledge for single aircraft model. The sources were: Aviation Safety Reporting System (Boeing-717 aircraft-specific entries), National Transportation Safety Board accident database, basic aviation information (including such documents as the Airman's Information Manual), Title 14 of the Code of Federal Regulations (CFR) Subchapter G, aviation weather, SKYbrary Aviation Safety, and Boeing-717 documentation (including the maintenance manual, Airplane Flight Manual, Flight Crew Operations Manual, and Quick Reference Handbook).

As part of the initial proof-of-concept, textual information was extracted from original source documents and the text was prepared for ingesting in to the WDA system. This preparation included removing non-English text, joining sentences which had been divided and splitting long documents (e.g., book-length documents were divided into chapters). In addition, charts, tables and other semi-structured materials were removed from the documents. (Note that ingesting this type of information is a crucial PEER Phase 3 enhancement.) The resulting document collection, consisting of 1,147 individual documents making up a corpus of more than 7.5 million words of text, was used as the basis of the domain adaptation experiments carried out in the current phase.

The process of ingesting documents into the WDA system involves various language processing and indexing steps which depend on the specifics of the domain lexicon, so whenever the lexicon was updated, the textual corpus was re-ingested as well. At the outset of the current project this corpus was ingested into the WDA system, making use of a minimal domain (177 term) lexicon developed in the Proof-of-Concept. Over the course of the project, there were three major lexicon updates. Details on the lexicon development are given below in the *Domain Lexicon Development* section.

Question-Answer Set Development

To both customize the ranking models and assess the effectiveness of domain adaptation activities, IBM and NASA collaborated to develop a set of domain-specific questions that had answers to be found in the NASA PEER document corpus. This "Question/Answer set" (QA-set) contained both questions specific to the Midwest Flight 490 use-case as well as questions about aviation and the aircraft in general. The QA-set, for training and testing WDA, is constrained to contain questions that are factual in nature and have a short, specific correct answer which can be explicitly listed. As the set plays the role of a "ground truth" in training and testing the WDA system, it is crucial that the WDA system be able to determine automatically whether or not an answer produced by the system was the correct answer. Examples of these questions are given below in Table 1.

Table 1: Question/Answer Examples

Question	Answer
What flap setting will initiate automatic go-around below 2500 feet?	5 degrees
Which cloud structure might lead to rapid ice build-up on the plane?	Cumuliform
Where on the plane is the CVR located?	Aft tail compartment
By what speed during the takeoff roll are anti-icing fluids shed?	100 knots
What is the frequency spacing of the VHF frequency bands?	25 MHz
What motion is caused by out of phase yaw and roll?	Dutch roll

The complete set of question/answer pairs generated for this project consisted of 508 questions of this sort, developed by the NASA SMEs and reviewed by the IBM team. Each question was checked to assure that the answer was contained in the document corpus, and that there was a single correct answer. The review process involved three rounds of quality assurance to determine that the questions were appropriate for training and testing the WDA system.

Once the final set of questions was reviewed and accepted, two subsets of the questions were distinguished, those to be used to train the system (the "training" set of 356) and those to be used to test the system (the "test" set of 152). Assignment of a question to the training or test set was random.

Domain Lexicon Development

A domain lexicon specifies what the domain terms are as well as what taxonomic relations hold among these terms. It is in the domain lexicon, for example, that we encode the information that "parasitic drag" is a kind of "drag" and that "drag" is a kind of "force" (along with "lift" and "thrust"). As suggested above, lexicons in WDA inform many aspects of the functionality of the system. The lexicon provides WDA with information about the relevance of terms in a passage to topics in the question, and it also informs the mechanisms that determine how well an answer matches a question. The WDA domain lexicon includes information about the taxonomic relations among terms (hyponymy or "is a type of") as well as information about domain synonyms (e.g. that in aviation "yoke" is another term for "column").

This kind of domain-specific lexicon is particularly important in technical domains such as aviation (or physics or medicine) because general purpose lexical resources such as WordNet (Fellbaum, 1998) or DBpedia (Auer et al, 2007) by their very nature often ignore specialized meanings of terms. For instance, in the aviation domain, the terms "lift" and "drag" have special meanings and should not be misconstrued with their more common variants, or with specialized meanings from other domains. That is, "lift" should be associated with the aerodynamic forces related to wings, velocity, altitude and thrust, not with picking up a box. Domain specific lexicons provide this information.

The challenge for lexicon development is the pace of development and the labor required for the hand-construction of domain lexicons. In the initial Proof-of-Concept, for example, a domain lexicon of just under 200 items was developed by Subject Matter Experts (SMEs) working with IBM engineers over the course of a four-week period. The goal for this project was to extend this domain lexicon with 500 lexical items. In the course of the project, it became clear that a larger scale lexicon than this was required. In the course of the data analysis of the document collection carried out prior to document ingestion, the IBM team identified a number of documents in the NASA PEER document collection that contained extensive domain glossaries. Recognizing that there was significant potential to extract lexical information directly from these glossaries in many of the documents, the IBM team looked to exploit these domain glossaries by automatically extracting lexicon information from these glossaries.

Lexicon Extraction from Domain Glossaries

The task of converting a glossary written for human consumption into a WDA domain lexicon essentially involves three processes: 1) extracting the glossary entries from the source text; 2) identifying equivalent terms and normalizing; and, 3) categorizing the entries for use by WDA.

The first stage of constructing the automatic lexicon was to generate a list of glossary entries from the text or manual and guidebooks. This step was highly dependent on the format and structure of the source documents. The source for this effort was plain text extracted from PDF versions of the manuals and guides. The entry extractor used cues in the text to find sections of the document to look for entries, such as the word "GLOSSARY" appearing by itself on a line. Within these sections, the extractor looked for a handful of patterns that were observed in the text that indicated a glossary entry.

Each entry in the glossary contains a term or phrase, an optional abbreviation or acronym, and some text, which is usually a definition. In some cases, the definition text included a cross reference (e.g. "See ...") or an acronym expansion as well as the full definition. Some examples of the observed formats follow:

> Corrected mean temperature (CMT). The average between the target temperature and the true air temperature at flight level.
>
> ALTIMETER—A flight instrument that indicates altitude by sensing pressure changes.
>
> Altimeter. A flight instrument that indicates the altitude above a given reference point.
>
> Alter heading (A/H). The change in heading to make good the intended course.
>
> COMPRESSOR BLEED AIR — See BLEED AIR.
>
> Declination (Dec). The angular distance to a body on the celestial sphere measured north or south through 90° from the celestial equator along the hour circle of the body. Comparable to latitude on the terrestrial sphere.
>
> Equinoctial. See Celestial Equator.

It was these kinds of expressions that served as the input to the automatic extraction system. The process of automatic glossary entry extraction yielded 3,592 glossary terms from 14 different domain files. These were extracted into a structured format identifying the term, an optional acronym and the gloss.

Additional processing was done to identify duplicate terms and to resolve cross references for acronyms. The resulting resource had the form illustrated in Table 2 below.

Table 2: Extracted Glossary Example

Term	*Acronym*	*Gloss*
Absolute pressure regulator		A valve used in a pneumatic system at the pump inlet to regulate the compressor inlet air pressure to prevent excessive speed variation and/or overspeeding of the compressor.
Absolute zero		The point at which all molecular motion ceases. Absolute zero is -460°F and -273°C.
Accelerate-Stop Distance Available	ASDA	The runway plus stopway length declared available and suitable for the acceleration and deceleration of an airplane aborting a takeoff.
Accelerate-go distance		The distance required to accelerate to V_1 with all engines at takeoff power, experience an engine failure at V_1, and continue the takeoff on the remaining engine(s). The runway required includes the distance required to climb to 35 feet by which time V_2 speed must be attained.

Categorizing glossary entries

With a set of entries extracted from the source text the next step is to categorize the entries – essentially inserting them into a taxonomy – so they are usable by WDA processes such as Answer Generation and Answer Ranking. This task included classifying terms – determining that “yoke” refers to a flight deck control, for example. This categorization was achieved by interpreting the text of the definition. This process leverages some well-established NLP approaches to categorize text and extract topical keywords from the entries (Velardi, Cucchiarelli, and Petit, 2007).

Clustering

Treating the gloss of a domain term as a document, we leveraged document clustering methods to cluster the glosses (and by extension, the domain terms they were associated with). There are a range of automatic document clustering technologies to accomplish this, and we experimented with a few variations before settling on K-means clustering (Hartigan and Wong, 1979). K-means provided an easy way to tune the clusters and also identified top cluster-words for each cluster. The automatically generated clusters were then mapped to existing WDA types (and to generate new sub-types of those types as well).

The K-means approach also gave us the ability to tune both how many terms were considered in the clusters and the number of clusters that were created. It turned out that the cluster quality, i.e. how accurate a cluster was in representing a specific topic or concept, was highly dependent on these two factors. Selecting too many or too few terms for the cluster training tended to place most entries in one large general cluster that had vague general terms like "flight" and "aircraft" as their key terms along with a handful of small clusters that represented highly specialized concepts like refueling nozzles.

After some experimentation the final clustering used the 1000 most relevant definition words and created 50 clusters. This produced coherent, reasonably sized clusters. The definition words for each cluster were used as a guide for a human reviewer to assign a cluster to a WDA type.

Headword Identification

A second approach was used in coordination with the automatic clustering to discover relevant categories for an entry. This approach was based on the intuition that most dictionary definitions begin with a statement of the type of information the entry is about. Examples include “Torque/A force that produces or tries to produce rotation”, “Elevator/The horizontal, movable primary control surface in the tail section”, and “Durability/A measure of engine life.” By parsing the definitions using a IBM dependency tree parser (McCord, Murdock, and Boguraev, 2012) and identifying the noun phrase headwords of the main noun of the definition (“force”, “control surface”, “measure”) as they key type term, the domain terms could be classified by these keys.

This process generated approximately 1,200 unique key type headword terms. A sample of the most frequent headwords extracted by this process is included in Table 3. The number of occurrences correspond to the number of glossary terms (domain terms) whose definition has the given headword. Note that the vast majority of the glossary terms had definitions which were unique (a single occurrence).

Table 3: Most Frequency Headwords

Headword	*Occurrences*
system	49
condition	47
device	46
line	37
distance	37
altitude	36
method	34
point	30
time	29
airspace	29
aircraft	29
pressure	24
instrument	24
ratio	22
weight	21
speed	21
angle	21
area	18
measure	16
force	16

Assigning Semantic Types

Mapping the headwords and clusters into a semantic taxonomy was a semi-automatic process. In a first phase, an IBM engineer reviewed the existing lexicon and associated an appropriate semantic category with the 50 automatically generated clusters and for the top 100 most-frequent extracted headwords. In almost all cases the headwords or cluster were associated with a novel domain-specific semantic type, such as cockpit display (“Control Display Unit”, “Synthetic Vision System”) or air spaces (“ATC

assigned", "Class D"). This manual assignment of keywords and headwords to semantic types induced an automatic classification of the novel terms, which were incorporated into the lexicon. The resulting lexicon was then reviewed, and minor errors adjusted. For example, the headword "approach" was associated both with terms related to the approach phase of a flight and with terms describing a methodology (e.g. "aeronautical decision making"). This kind of disambiguation was accomplished through manual review.

Manual Extensions

Beyond the large-scale automatic domain adaptation effort described above, the lexicon was extended in two other ways: a) by exploiting the implicit lexical information in the question-answer training set; and, b) by exploiting external domain-specific resources.

For the 356 training-set QA pairs, an additional hand-check was done to identify missing lexicon material and to determine the effectiveness of the automatic extraction. The answer key was used to identify lexical entries crucial for question answering. For example, if a question targets a domain-specific semantic type (as "flight instrument" is targeted in "Which flight instrument shows how high the aircraft is?"), it is crucial that the answer term (e.g. "altimeter") be included in the domain lexicon. While this does not exhaust the use of domain lexicons in the WDA system, it is a good heuristic for coverage. For over 250 of the questions, automatic domain adaptation had identified the crucial lexical entries. For the remaining questions an alignment check identified 64 additional lexical entries to be added to the lexicon, filling in terminology gaps.

For certain semantic types, it was found to be advantageous to exploit even more highly domain specific lexical resources than the domain glossaries. In particular, three aviation-specific and aircraft specific lists were identified by IBM consultants and used as the basis for lexicon expansion. This includes the following available external resources:

List of CFR chapters	https://www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title14/14tab_02.tpl
Flight phase taxonomy	https://www.skybrary.aero/index.php/Flight_Phase_Taxonomy
List of 717 alerts	http://www.gabertech.com/B717cbt/

The final domain lexicon contained just over 4,500 terms.

Model Training

The final component of domain adaptation for WDA involves training machine learning models. This aspect of domain adaptation tunes the answer-ranking models to the domain-specific source data (the corpus) as well as the domain specific question set and the domain lexicon. For this step, the 356-question training set was leveraged. This set was small with respect to typical training sets for WDA (which range in the thousands of QA pairs), therefore initial training and testing was done to evaluate the effectiveness of training the machine learning models on the basis of only the small set of domain questions. On the basis of this result, the domain-specific training set was augmented with domain-general training, using domain-general source data (Wikipedia) alongside the domain-specific lexicon. Training proceeded in two phases, the first phase leveraged the initial automatically extracted lexicon, and the second phase included the additional manual extensions of the lexicon.

Other Domain Adaptation

While lexicon-based domain adaptation is the primary source of customization for the WDA system, two other mechanisms were available for customization: one related to acronyms and the other related to numerical questions.

Additional configuration of the WDA system involved the exploitation of domain-specific acronym expansions. Acronym expansion is used in the WDA system as a resource for query expansion. An acronym expansion resource consisting of nearly 1000 aviation domain specific acronyms was generated and used to configure the PEER system. These acronyms were extracted from the NASA PEER corpus using targeted information-extraction techniques identifying acronyms and expansions in the ingested documents. The query extraction mechanism was configured to use these acronyms.

The WDA system also has specific mechanisms for identifying questions with numerical answers, which can be customized to the domain. In the aviation domain, of course, terms such as “angle”, “load factor” and “RVR” are indicators of numerical answers. The training QA set was analyzed to extract a set of domain specific numerical indicators and the numerical answers recognizer configured with these.

These activities constituted the extent of the domain adaptation of the WDA system to the PEER use case.

Error Analysis and Assessment

Detailed analysis of the system on the basis of its observed performance is an important indication of the potential for system improvement. As part of this project, IBM performed error analysis to assess the kinds of issues that the aviation domain posed for the WDA question-answering pipeline.

Error Analysis

Both prior to and after lexical domain adaptation, configuration and system training, an error analysis was undertaken on the test-set of questions to identify sources of error for potential system improvement. Error analysis can uncover areas of domain adaptation that have been neglected as well as identifying other aspects of the configuration which were sources of the system's failure to find the correct answer to a test question.

As part of error analysis, test questions were posed to the system, and on the basis of the answers returned the questions were classified into one of the following five classes:

- **No Search Hits**: Retrieval mechanism did not find passage with the correct answer in it
- **Unextracted Candidates**: Retrieval mechanism found passages with answer in it, but answer was not proposed as candidate
- **Imperfect Answers**: Retrieval mechanism found passage with the correct answer in it, but the answers that were proposed only partially match the answer key
- **Exact in Ranks 11-100:** Highest ranked correct answer was between ranks 100 and 11
- **Exact in Ranks 6-10:** Highest ranked correct answer was between rank 6 and 10
- **Exact in Ranks 2-5:** Highest ranked correct answer was between rank 2 and 5
- **Exact in Rank 1**: Highest ranked correct answer

This classification indicates roughly the source of the error, be it the Corpus Search, Answer Generation or Answer ranking. The initial round of error analysis identified that a significant number of the *No Search Hits* questions contained acronyms, and the acronym expansion model which was used as part of the retrieval mechanism was, therefore, more narrowly focused on the NASA PEER acronym expansion. In the preliminary stage it was clear that the predominant source of error was in the Answer Generation component of the system. As illustrated in Figure 4 below we see the large proportion of *Unextracted Candidates* – questions for which the correct answer is found in the passages retrieved, but which the Answer Generation component did not propose as an answer. Initially, Answer Ranking is not a major source of errors – when the correct answer is generated, it is mostly ranked at least in the top five answers. Subsequent to domain adaptation this changed, however, and the predominant source of errors was the Answer Ranking components – more correct answers were generated but ranked lower.

In other words, subsequent to domain adaptation, the system proposed the correct answer to most of the questions, but that answer was not ranked highly enough. This points to an anticipated weakness in the machine learning ranking models, since for this project the set of questions used to train the ranking models was quite small. The PEER prototype system was trained with only 356 questions, while the Watson Jeopardy! system was trained with upwards of 50,000 questions. Significant improvements are anticipated with even moderately more training data.

This quantitative Error Analysis informs the more qualitative analysis of the kinds of errors that the WDA system is making. Among the most pronounced systematic errors made by the system were the errors made in relation to location questions. WDA is designed to answer questions about geographical locations (such as "Where was the first president of the United States born?") but not about the non-geographical locations. Questions such as "Where are the engine fire shutoff valves?" which are clearly central to the PEER use case turn out to be extremely problematic for the system. (Parallel issues arise in the context

non-calendric temporal questions such as “When should nose up elevator be applied?”) A second type of question that proved to be problematic are those whose question focus was non-sortal (Guarino 1995), meaning that the focus of the question was a term such as “malfunction” or “condition” that is highly contextual in its meaning. WDA’s main processing question-answering heuristics are tied to identifying the ontological type of the entity sought as the answer to the question. This is effective for questions with sortal focus, such as “What city hosted the 1960 Winter Olympics?” It works less well for questions such as “What condition does prolonged use of electronic equipment on the ground cause when the avionics fan is inoperable on the ground?” in which the answer (“overheating”) does not have a well-defined ontological type. This error analysis points to enhancements of the WDA question-answering pipeline to address non-geographic “where” (and non-calendric “when”) questions as well as non-sortal focus questions.

Figure 4: Error Classification for Baseline System

Assessment

The Watson Discovery Advisor system provides answers to naturally posed questions in a domain, calculating for each potential answer to a question, the system’s confidence that this is the answer. One way of assessing the Watson system pioneered by the IBM research team in developing the Watson Jeopardy! system is to ask how many questions are correct at given thresholds of answering. As

illustrated in Figure 5, at different thresholds for answering, different values for precision are observed (Ferrucci et al, 2010).

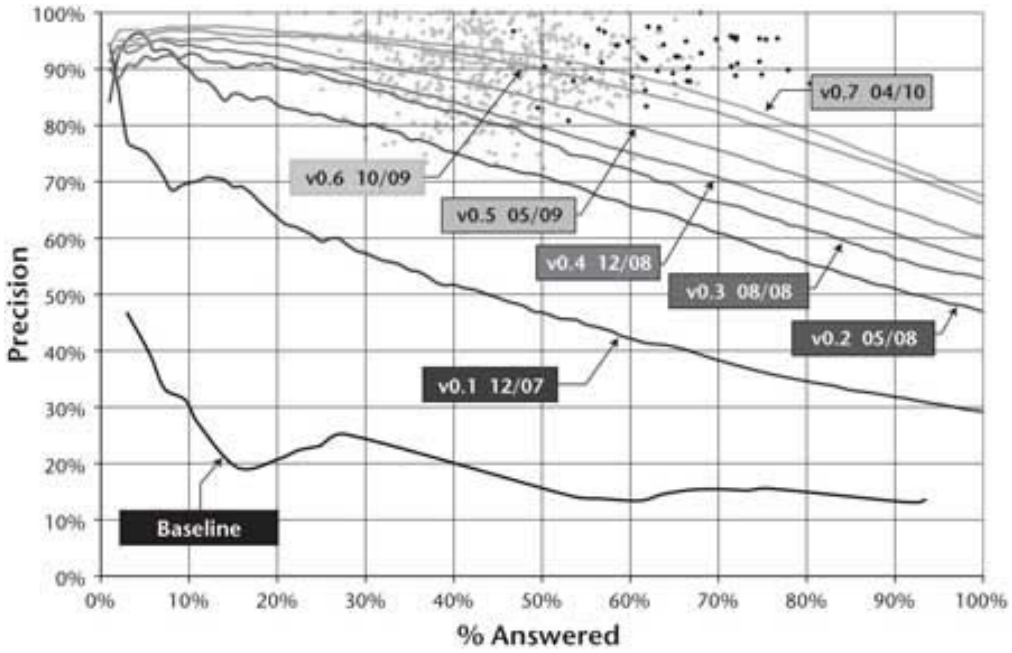


Figure 5: Watson Jeopardy! System Improvements Over Time

This graph illustrates an assessment of the correctness ("precision") of the system for different subsets of the test set. The subsets are arrayed in terms of the confidence the system has in the answer. The points corresponding to 50% answered indicate the assessment of the system performance on the half of the test questions that the system is most confident in. In the initial baseline system, for example, when the 50% of the questions about which the system was most confident – the "top" half - were considered, the system got about 15% of them correct. When considering the top 10% of the question, however, the system answered 30% of the questions correctly. This representation gives a sense of how effectively the system is answering questions as well as how well it is assessing its own confidence in the answer. As we would expect, by and large the more confident the system is, the more likely it is to be correct. Over the course of four years of development, we see the curve slowly rise across the entire range.

For the PEER use case, the system being designed is expected to deliver only high confidence, high accuracy information. Since we don't yet know, however, what an appropriate threshold would be for returning an answer – that is an issue for further research – we provide an assessment of the value of domain adaptation across the whole range of answered-questions thresholds. Results are displayed in Figures 6, 7, and 8 below.

In the three figures below, we also present additional information about whether the correct answer is within the top 5-ranked (blue line) and top 3 ranked (orange line) as well as the top-ranked answer to each question considered. This provides insight into how "close" the system was to ranking the correct answer as the top answer. Using the 508 question QA set collected in the first phase of this project, we assessed the WDA system by comparing the system response prior to domain adaptation and configuration with the system response after domain adaptation and configuration had been carried out. Assessment was done both with respect to a blind set (152 questions which were set aside and not used in any part of the domain adaptation) and with respect to the full set.

In Figure 6, the response of the pre-adaptation ("baseline") WDA PEER system is illustrated. As expected, absent significant domain adaptation, the WDA system's response is moderately effective, with

about a 30% of the questions answered correctly for the most confident tenth and one-fifth of the questions answered correctly for the most confident quarter. Note that in the most restrictive range on the left, low-number effects start to evidence themselves.

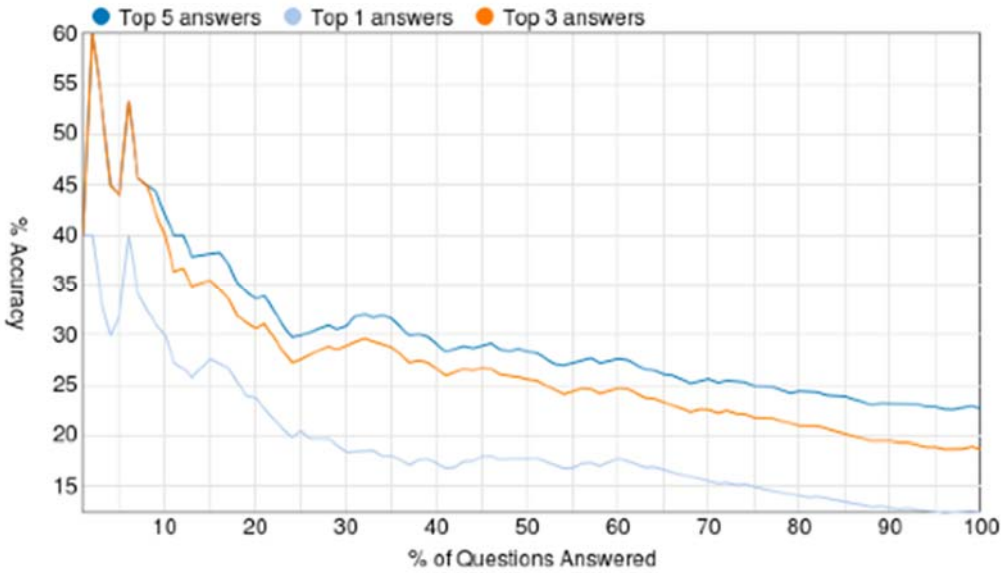


Figure 6: WDA PEER, Proof-of-Concept Baseline

In Figure 7, the assessment of the domain-adapted WDA PEER system as applied to the blind test set is shown. As we see, the quality of the system response is elevated across the board. In the most confident 10% about 40% of the questions are answered correctly. Across essentially the entire spectrum of questions, the domain-adapted system is more accurate than the baseline – in answering the most confident half of the questions the baseline system got about 17% correct, while the domain adapted system got about 25% correct. The comparison for the most confident 10% shows a similar improvement due to domain adaptation.

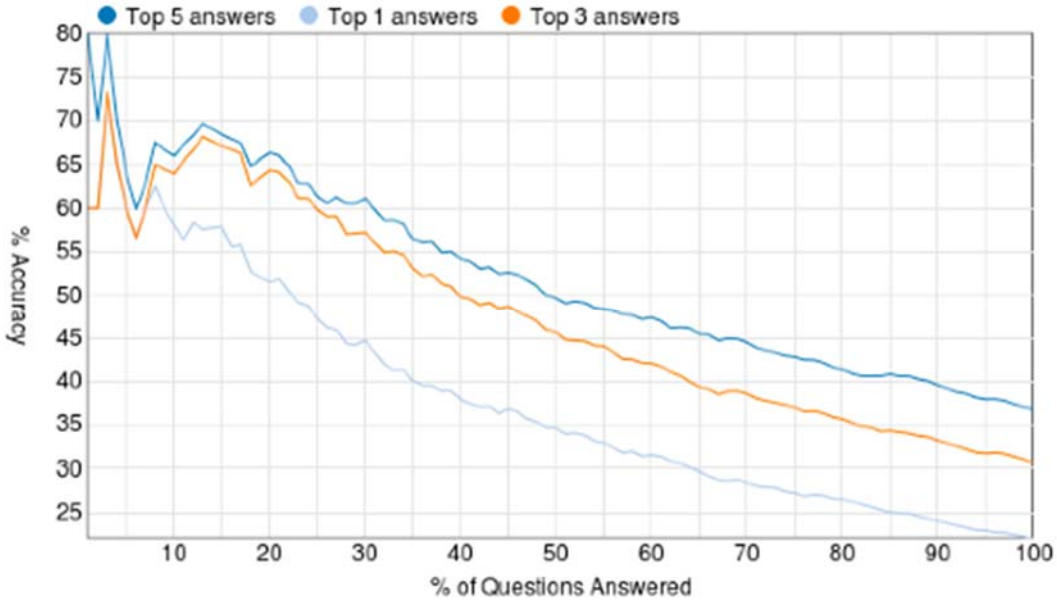


Figure 7: WDA PEER, Domain-Adapted, blind test assessment

Finally, in Figure 8, the assessment of the WDA system as applied to the entire set of 508 items is displayed. This figure illustrates the degree to which target domain adaptation dramatically improves the quality of the most confident answers. What is notable about this question set is that the domain lexicon coverage – the degree to which the terms in the question are to be found in the domain lexicon – is extremely high.

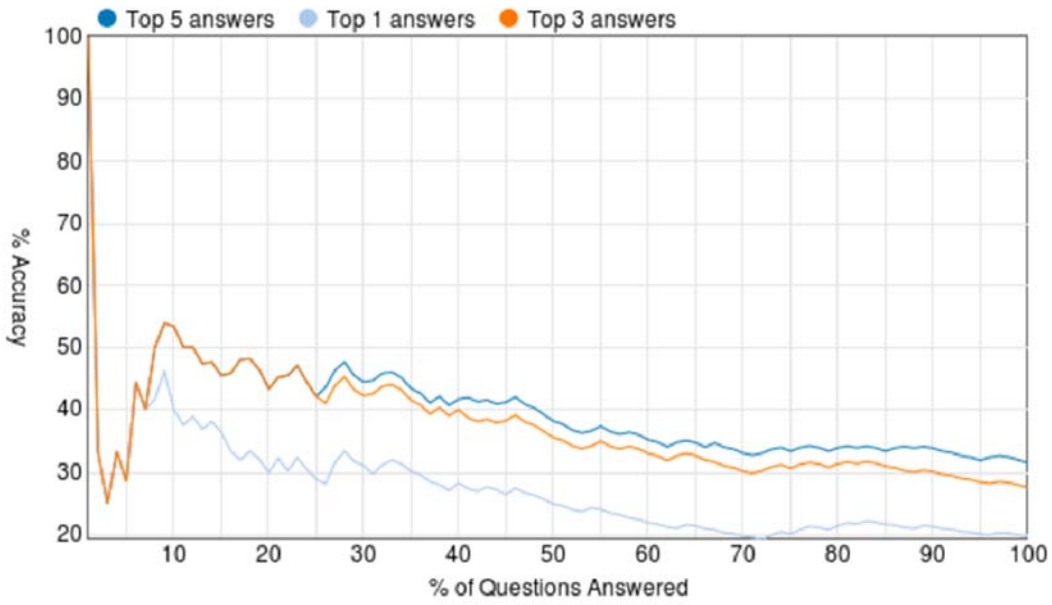


Figure 8: WDA PEER, Domain-Adapted, Full Assessment

This figure is indicative of the kind results that might be expected from a system that has a more complete domain lexicon, because a large proportion of the question in the full assessment set had been subjected to specific targeted domain adaptation.

Contextual Data Augmentation Design

Contextual information, including information about the state of the aircraft and of aircraft systems, the nature of the current flight plan, the air traffic situation, what the weather is like and so on plays a central role in identifying information relevant to flight crews. For the Pilot-Engaged Expert Response system to generate accurate and contextually relevant responses to naturally formulated queries from the flight crew, the system requires a mechanism for identifying contextual information relevant to a query, accessing that contextual information, and leveraging that contextual information in generating system responses. For example, to answer the question "Where is the fire extinguisher in the passenger cabin?" the crucial contextual information about the type of aircraft should augment the question – effectively sending the question "Where is the fire extinguisher in the passenger cabin of the B-717?" to the system. The design for a subsystem of the PEER system that exploits contextual information to improve the usefulness and accuracy of the system was a second goal for this project.

The Contextual Data Augmentation (CDA) subsystem of the PEER system is designed to have three main components: Contextual Data Collection, Contextual Data Selection and Contextual QA Augmentation. These components are organized into a processing pipeline: Once a query is posed to the system, PEER will start the Contextual Data Collection process. Then the Contextual Data Selection process will apply to select for augmentation the most relevant contextual data. Then the Contextual QA Augmentation process will leverage the selected contextual data for question answer, augmenting the question answering process with this information. In the CDA subsystem, two (2) data stores are used to store both historical data from all contextual sources (Historical Data Store) and operational data for a query (PEER Data Store). The functional architecture is illustrated in Figure 9.

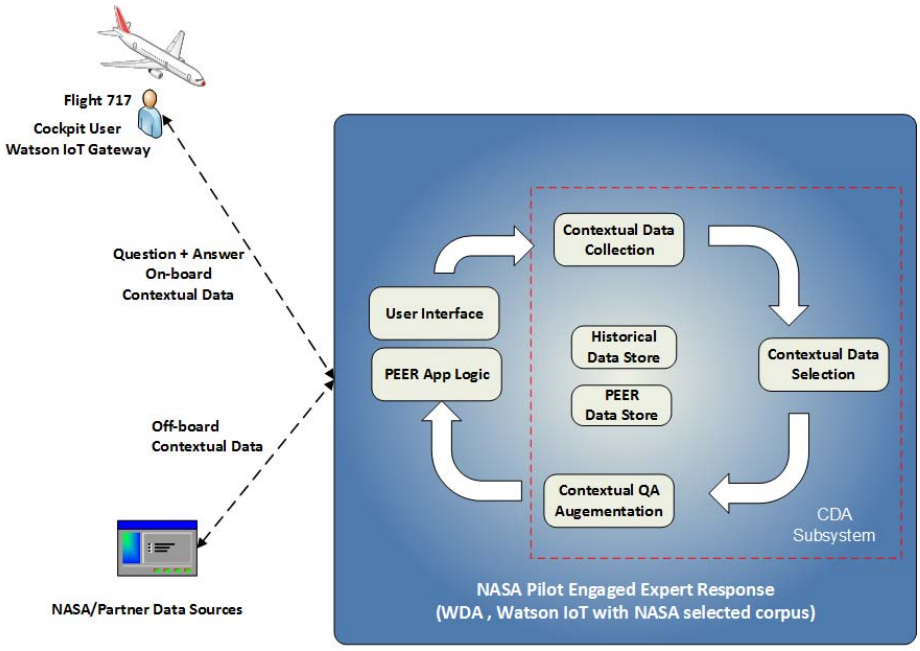


Figure 9: PEER Functional Overview

As a simple illustration, the steps in processing a query such as "What flight systems need to be checked?" as posed to the PEER system at 10 PM while the aircraft is on approach are shown in Table 4.

Table 4: Steps in Contextual Data Augmentation

Question: "What flight systems need to be checked?"
Step 1: Contextual Data Collection: Assembles information about current context (<Time-of-Day: *10 pm*, Altitude: *5000* ft, Flight-Stage: *On Approach*, Alerts: NONE, …>)
Step 2: Contextual Data Selection: Determines that Time-of-Day, Flight-Stage and Alert information could be relevant to current question
Step 3: Contextual QA Augmentation: Augments QA processing with Time-of-Day and Flight-Stage information (no cockpit alerts showing) to build enhanced question: "What flight systems need to be checked?" Context: "at night"; "during landing"

Relevant context data can augment the response-generation process in a number of ways; in addition to the simple augmentation of the query, contextual data can be used within the Watson QA pipeline to contribute to specific components such as answer scoring and response filtering. The following sections will describe the three key functional modules of Context Data Collection, Contextual Data Selection and QA Augmentation in more detail.

Contextual Data Collection

Contextual Data Collection is the collection of relevant data for potential use in augmenting an informational question to the PEER system. The data collection mechanisms will store this information in a central repository of contextual data for use in QA augmentation. As part of this processing and data collection, PEER will use an Internet of Things (IoT) Gateway to bring edge analytics to the cockpit so that collected data from the aircraft data bus structure can be analyzed and responded to locally without consuming much data bandwidth which is precious on the aircraft. It leverages a streaming engine which is optimized for edge processing.

A modern aircraft can include somewhere around 20K to 400K data points, so the capability of cockpit level preprocessing will be essential to create a practical solution. Only crucial contextual data will be selected and forwarded to the IoT Platform and WDA on the ground for: a) additional analytic processing; b) PEER contextual question answering; and, c) storage in a historical repository for future analysis. The kind of additional analytic processing envisioned as part of the data collection process includes processes needed to identify features of the dynamic contextual data derived from time-series analysis, such as rates-of-change of readings, or other crucial dynamic properties of indicators, such as the identification of most-recent changes in the cockpit control panel state to identify new alerts.

Functionally, the Contextual Data Collection component makes available the information about the current context (and past context) for use in QA augmentation. It stores associated contextual parameter features with their values, as illustrated conceptually below in Table 5.

Table 5: Contextual Data Store (Conceptual)

Context	*Contextual Data Values*
Current Context	*Flight_Stage*: Landing; *Altitude*: 1K ft; *Time_of_Day*: 10 PM, …
Past Context -1	*Flight_Stage*: Cruise; *Altitude*: 20K ft; *Time_of_Day*: 9 PM …
Past Context -2	*Flight_Stage*: Takeoff; *Altitude*: 1K ft; *Time_of_Day*: 8 PM, …

This store of contextual data (particularly the values of the current context store) are used by the other modules of the CDA subsystems.

Contextual Data Selection

There is a wide range of data that could, in principle, provide context for a given question and be used to augment that question in the Discovery process. Only certain contextual information will be relevant to a specific question, however; information about altitude will not be relevant to questions about the location of the fire extinguisher (but information about aircraft type will be). An important step to augmenting a question with contextual information is to select which data is relevant to the question at hand. Determining which contextual data is to be used is the task of the Contextual Data Selection sub-component of the CDA system. This component specifies for any question, what information will be used to augment that question. The problem of contextual data selection is divided into two parts: the training of a Contextual Data Selection model and the application of that model to the run-time processing.

The Contextual Data Selection model is a model that determines, for a given question, what types of information are likely to be relevant. For example, for the question: "Where is the fire extinguisher in the passenger cabin?" contextual information about the Aircraft-Type is likely to be relevant, while for the question "What does that alert mean?" information about the Most-Recent-Alert is likely to be relevant. To identify which information is relevant for a particular question, the following three step type-based method is proposed for training a data selection model:

1. Contextual data classified according to Contextual Data Type (CDT)
2. Training Questions labeled with the CDTs relevant to them
3. A Contextual Data Type model is trained to assign CDTs to unseen queries

While the contextual data that will be used to augment a specific question are specific to the circumstances of that question, such as that the current Aircraft-Type is a "Boeing 717" or that the Most-Recent-Alert is a "RUDDER-LIM-FAIL", the determination about relevance will be done on the basis of the type-based model.

To this end, each contextual feature is to be classified by type. Some of this classification being done by the Contextual Data Collection analytic component. Illustrative examples of Contextual Feature Types are given in Table 6.

Table 6: Contextual Features Types (conceptual)

Time-of-Day	Flight-Stage	Temperature	Aircraft	Weather
7 AM	Taxi	90°	Boeing 717	Clear
2 PM	Takeoff	75°	Airbus A380	Foggy
5 PM	Cruise	40°	Boeing 747	Thunderstorms
10 PM	Landing	20°		

For the purposes of the CDA system, the specific context of a question is represented as a vector of values for each of these contextual features at the time that the question is posed. For example, a question made in on an Airbus landing on a foggy hot night could have context vector represented as:

<*Time-of-Day*: 10 PM, *Flight Stage*: Landing, *Temperature*: 90°, *Aircraft*: Airbus A380, *Weather*: Foggy>

Real-world context vectors for PEER will encompass dozens if not hundreds of features, most of which would be entirely irrelevant to a given question. The problem for the contextual selection model is to determine for any particular question what elements of the context might contribute to the Discovery process for that question. It is expected that there will be extreme variation in both the number of types of contextual data used to augment a question as well as the types of information. In highly contextual

queries, such as “What do I do now?” it is expected that many types of contextual information will be relevant, whereas in more narrowly drawn queries, such as “What systems have to be turned on prior to landing?” perhaps only a few contextual features are important.

The determination of which contextual data is relevant to a given question is dependent upon a number of factors, including the corpus of textual data in the Discovery component as well as the specific topic of the question posed to the system and the value of the contextual parameter. As a practical matter, a piece of context can be determined to be relevant to a question if that question is more effectively answered when it is augmented with that piece of context. It is this data that will be leveraged in training the Contextual Data Selection model.

The Contextual Data Selection model will identify what kind of contextual features are relevant for any given question in a given context. In the most general case, the input space would be the Cartesian product of the questions and the context. It is anticipated that two models for identifying relevant contextual features will be trained: A context independent model, which is trained to label questions with contextual parameter types irrespective of the particular contextual value they have; and a context dependent model, in which the values of the contextual parameters determine whether the feature is selected. The context independent model will be used to label questions with the relevant features. The context dependent model will be used to identify features of general relevance. In deployment, each of these models will be run to derive the set of contextual factors to consider for a given a question. For example, in our illustrative case, for a question (Q), the following contextual data selection (C) is proposed:

> Q: “What systems have to be turned on?”
> C: <*Flight_Stage*: Takeoff; *Altitude*: 5K; *Time_of_Day*: 10 AM, …>

Context Independent Selection would apply to (Q) to give, perhaps, *Flight_Stage*; *Time_of_Day,* etc. because these parameters are relevant to questions like this and the Context Dependent Selection would apply to (C) to give, perhaps *Altitude*, because this parameter is relevant because of its value. Training these selection models will be crucial to the success of the Contextual Data Augmentation module.

Contextual QA Augmentation

Once the context data are selected, they will be used to augment the WDA question-answering. This involves making use of contextual information the components in the QA pipeline outlined above. The components of the QA pipeline that will be the focus of contextual augmentation are the Corpus Query and the Answer Ranking component. For Corpus Query, QA augmentation will typically involve adding contextual information directly to the query via a Query augmentation module. For Answer Ranking, QA augmentation will play a role in determining the goodness of fit of the answer.

To illustrate some of these components working consider the conceptual example: “What systems have to be turned on?” For this example, the Context Selector component generates the list of contextual data that will be used in the QA pipeline to augment this question to the system. For example, Context Selector might return <*Flight_Stage, Altitude, Time_of_Day*>.

Once data is selected, the augmentation system component is responsible for retrieving the values (e.g. ‘5000 feet’) of selected contextual data (e.g. ‘Altitude’) from the Historical Data Store into the PEER Data Store for QA operational use. The PEER application logic will retrieve the values from the PEER Data Store of the contextual parameters for the current context, for example:

> **Question**: “What systems have to be turned on?”
> **Context**: <*Flight_Stage*: Landing, *Altitude*: 5K ft, *Time_of_Day*: 10 PM>

Once the contextual features are fully specified, QA augmentation takes place. The Query Builder component will use the context to produce a query from the contextually augmented question structure.

> **Question**: "What systems have to be turned on?"
> **Context**: <*Flight_Stage*: Landing, *Altitude*: 5K ft, *Time_of_Day*: 10 PM>
> **Query**: +contents:system +contents:turn +contents:10pm spanNear([contents:turn, contents:on], 2, false)^0.4 spanNear([contents:take, contents:off], 2, false)^0.4 spanNear([contents:fly, contents:off], 2, false)^0.4 spanNear([contents:above, contents:ground], 2, false)^0.4 contents:5000 contents:5K

Finally, Answer Ranking will use the contextual information and provide the answer rankings back to PEER based on a specific scoring algorithm. This answer score is part of the context dependent scorer framework currently supported in WDA. We illustrate this conceptually below (actual scores will depend on the configuration of the system).

> **Question**: "What systems have to be turned on?"
> **Context**: <*Flight_Stage*: Landing, *Altitude*: 5K ft, *Time_of_Day*: 10 PM>
> **Query**: +contents:system +contents:turn +contents:night spanNear([contents:turn, contents:on], 2, false)^0.4 spanNear([contents:take, contents:off], 2, false)^0.4 spanNear([contents:fly, contents:off], 2, false)^0.4 spanNear([contents:above, contents:ground], 2, false)^0.4 contents:5000 contents:5K
> **Passage**: "Lighting systems should be turned on at night when landing"
> **Answer**: "lighting systems"; ContextPASscorer:0.5; ContextTyCor: 0.6

Conclusion and Next Steps

The current project represents the initial phase of development of the Pilot Engaged Expert Response (PEER) system based on the Watson Discovery Advisor system. This first step realizes part of the vision for PEER as an information resource for flight crew, providing access to a wide range of knowledge about the aviation domain from a wide range of domain documents. This project has focused on the knowledge retrieval and storage components of the system. In this project, an instance of WDA was successfully adapted to the aviation domain through domain lexicon development and domain specific training, enabling users to ask questions about aviation topics and receive useful and accurate answers to these questions. Procedures for adapting the system to a technical domain through automatic lexicon extraction from domain glossaries have been developed and refined, and these lexicons have been shown to dramatically improve the ability of the WDA system to answer domain-relevant questions. Question-answer training data were assembled and used to train the system and to demonstrate a dramatic improvement in the ability of the WDA system to answer domain-relevant questions facilitated by this domain adaptation.

In addition, the vision for the PEER system has been pushed forward by the articulation of a plan for the automatic enhancement of question-answering with contextual information that includes a sophisticated three component module for contextual data augmentation. This contains data collection, data selection, and query augmentation sub-components.

The next steps in the development of the PEER system encompass three areas of development: First, the implementation of the contextual question answering augmentation module designed in the course of this project; Second, discovery enhancements needed to address issues raised by the PEER QA set the incorporations of enhancements to the WDA core that are responsive to specific needs identified in the course of the project; Third, moving beyond the WDA interface through the initial development of the aircraft adapter modules components for communicating with a (perhaps simulated) aircraft and flight crew. Each of these tasks was specifically identified as filling technological gaps identified or further researched in the current project. Specific recommendations include the following:

i. Building of a Contextual Data Augmentation proof of concept, focusing on static data**,** simple hand-built models for contextual data selection, with automatic context sensitive query generation and context sensitive answer scoring,
ii. Development of domain-specific answer generation and scoring for problematic question types such as non-geographical location questions and non-sortal focus questions, and
iii. Building an initial response-filtering component, making use of confidence thresholds, and perhaps the development of a schema-based automatic question generation module based on flight deck alert.

In addition to these tasks identified in the current project, the incorporation and further development of technologies for deriving answers from tables and other semi-structured textual sources remains a crucial area for enhancement. So much of the knowledge is stored in technical aviation documents, that it is crucial that the latest IBM enhancements in the area of question answering from semi-structured data be applied to the PEER document corpus.

It is clear from the current project, that the PEER vision of an automated knowledge assistant available to flight crews is a realistic goal, and that incremental development of the capability is a way to make this goal realizable.

Bibliography

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. 2007. "DBpedia: A Nucleus For A Web Of Open Data." *The Semantic Web,* pp. 722-735.

Beller, C., Katz, G., Ginsberg, A., Phipps, C., Bethard, S., Chase, P., Shek, E., and Summers, K. 2016. "Watson Discovery Advisor: Question-Answering In An Industrial Setting." *Proceedings of the Workshop on Human-Computer Question Answering.* pp. 1-7.

Fellbaum, Christiane. 1998. *WordNet.* John Wiley & Sons, Inc.

Ferrucci D., Brown E., Chu-Carroll J., Fan J., Gondek D., Kalyanpur A., Lally A., Murdock J., Hyberg E., Prager J., and Schlaerfer N. 2010. "The AI Behind Watson—The Technical Article." *The AI Magazine.*

Ferrucci, David. 2012. "Introduction to "This is Watson"." *IBM Journal of Research and Development.* Vol. 56, No. 3.4.

Guarino, Nicola. 1995. "Formal Ontology, Conceptual Analysis And Knowledge Representation." *International Journal Of Human-Computer Studies.* pp. 625-640.

Harford, Tim. 2016. "Crash: How Computers Are Setting Us Up For Disaster." *Guardian*, Oct 2016.

Hartigan, J.A., and Wong, M.A. 1979. "Algorithm AS 136: A k-means Clustering Algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics).* pp. 100-108.

McCord, M.C., Murdock, J.W., and Boguraev, B. 2012. "Deep Parsing in Watson." *IBM Journal of Research and Development.*

National Transportation Safety Board. 2018. "NTSB Identification: NYC05MA083." *https://www.ntsb.gov/_layouts/ntsb.aviation/brief2.aspx?ev_id=20050523X00653&ntsbno=NYC05MA083&akey=1.*

Velardi, P., Cucchiarelli, A., and Petit, M. 2007. "A Taxonomy Learning Method And Its Application To Characterize A Scientific Web Community." *IEEE Transactions on Knowledge and Data Engineering.* Vol. 19, No. 2.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01/10/2018	CONTRACTOR REPORT	6/28/2017 - 1/31/2018

4. TITLE AND SUBTITLE

NASA Pilot Engaged Expert Response using IBM Watson Technology - Prototype Evaluation of Knowledge Retrieval System Final Report

5a. CONTRACT NUMBER
NNG15SC15B

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

Katz, Graham; Ding, Chengmin; Doyle Andrew

5d. PROJECT NUMBER

5e. TASK NUMBER
5.0.B

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

IBM COPORATION
6710 ROCKLEDGE DRIVE
BETHESDA MD 20817-1834

8. PERFORMING ORGANIZATION REPORT NUMBER
Order Number: 80LARC17F0005

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
NASA Langley Research Center
1 Nasa Drive
Hampton, VA 23666

10. SPONSOR/MONITOR'S ACRONYM(S)
NASA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
NASA-CR-2018-220097

12. DISTRIBUTION / AVAILABILITY STATEMENT
Unclassified - Unlimited
Subject Category 82
Availability: NASA STI Program (757) 864-9658

13. SUPPLEMENTARY NOTES
Langley Technical Monitor: Randall E. Bailey

14. ABSTRACT

15. SUBJECT TERMS
Watson; Artificial intelligence; Big data; Cognitive computing; Data; Data analysis; Data science; Machine intelligence; Machine learning; Watson Discovery Advisor

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
Unclassified - Unlimited					STI Help Desk (email: help@sti.nasa.gov)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U	U	35	(757) 864-9658

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18