Cognitive Assistance in Government
Papers from the AAAI 2015 Fall Symposium

Machine Interface for Contracting Assistance

Jason E. Summers, Daniel T. Redmond, Charles F. Gaumond
Applied Research in Acoustics LLC
1222 4th Street SW, Washington, DC 20024-2302
jason.e.summers@ariacoustics.com, daniel.t.redmond@ariacoustics.com, charles.f.gaumond@ariacoustics.com

Abstract

We describe a cognitive assistant in early-stage development for the United States Air Force as an aid to contracting officers and potential commercial offerors for navigating the government-contracting process. The goal is easing compliance and affording flexibility and transparency so as to support an innovative and rapid acquisition process. The motivation, use cases, and technical approach for MICA, a Machine Interface for Contracting Assistance, are discussed here along with the technical challenges posed.

Introduction

The Need for Contracting Assistance

The Defense Science Board (2008) has indicated that "U.S. Government policies, practices, and processes do not facilitate the development, deployment, and support of the innovative, affordable, and rapidly acquired weapons, systems, and services needed for the 21st century forces." Complexity of government-acquisition regulations has limited participation by those very nontraditional providers able to support innovative, affordable, and rapid acquisition by creating barriers to entry and unfavorably shifting the risk/reward curve. Understanding requirements—or even acquisition needs—requires investment of a significant amount of labor, which has significant opportunity costs for small businesses. Putting required programs and policies in place—such as Defense Contract Audit Agency (DCAA) compliant accounting, Defense Contract Management Agency (DCMA) compliant purchasing systems, or even required awareness training on combating trafficking in persons (FAR 52.222-50, 22 U.S.C. 7104(g))—imposes additional, sometimes significant, recurring fixed costs for labor, systems, software, and outside services (viz., accounting and legal). Likewise, negotiating and understanding government rights in intellectual property (IP) that vary across agency and contract vehicle and can change as a result of new legislation (cf., changes legislated in Section 824, FY 2011 National Defense Authorization Act), further biases the acquisition process toward risk and away from reward.

Proposal: A Cognitive Assistant

We suggest that these challenges can be addressed, in part, through a cognitive assistant, MICA, a Machine Interface for Contracting Assistance, which aids users in navigating and maintaining compliance with the requirements of government-acquisition regulations, while affording flexibility and transparency in the process.

Currently under early-stage development for the United States Air Force, MICA is a cognitive assistant that provides users from both government and current and prospective contractor organizations insight and clarification into what is required of them by defense-contracting statutes, regulations, practices, and policies through a natural-language question-and-answer system.

Beyond providing a natural-language query interface to defense-contracting statutes, regulations, practices, and policies such as the Federal Acquisition Regulations (FAR) and Defense Federal Acquisition Regulation Supplement (DFARS), the MICA system will leverage the (semantic) database of government requirements to automatically determine requirements for solicitations (e.g., those listed on FedBizOpps and accessed through the FBOpen API) and provide those requirements to potential offerors in a natural-language format that supports user-driven inquiry and drill-down.

This capability for automated generation of requirements associated with specific solicitations naturally extends to automated specification of required offeror capabilities, which further increases accessibility to nontraditional defense contractors and aligns with the continuing Department of Defense (DoD) transformation to a capabilities based, net-centric acquisition approach.

Use Cases

The functionality and technical requirements of MICA are best elucidated through use cases and explicit examples of those use cases. While not all proposed use cases involve a question-and-answer paradigm, it is helpful to introduce a taxonomy of questions (after Peter Clark), that classifies possible queries into twelve types and two categories. The first category is *simple questions* comprising

1. True/false
    - *"Do cost-type contracts require DCAA approval?"*
2. Find value/values
    - *"What is the simplified acquisition threshold?"*
    - *"What are the clauses required in a non-CAS CPFF contract?"*
3. Subsumption
    - *"Are all time-and-materials contracts cost-type contracts?"*
4. Cardinality
    - *"What is the number of clauses in FAR that provide exemptions to full-and-open competition?"*
5. Taxonomic
    - *"What kinds of contracts are there?"*
    - *"What category of thing is an ACRN?"*
6. Possibility
    - *"Can a business that is majority owned by venture-capital firms receive SBIR funding under a Phase II award?"*
7. Meta-Reasoning
    a. Functional dependence
        - *"Does the total value of the contract affect what clauses must be included?"*
    b. Following from premise
        - *"The total value of the contract is below the simplified acquisition threshold, does this mean the contract is exempt from mandatory VETS-100A employment reports on veterans?"*

These questions require answers that can be coded numerically (binary yes/no, integer cardinality, or real-number values) or in terms of the entries in or structure of a knowledge base (category or member names).

The second category is *complex questions* comprising

8. Definition
   "What is required anti-trafficking in persons training?"
9. Description
   "Describe the actions and steps that must be taken during a DCAA pre-award audit to determine adequacy."
10. Example
   "What is an example of a sole-source justification?"
11. Similarity
   "What are the differences between CAS and non-CAS accounting requirements for a cost-type contract?"
12. Relationship
   "What is the relationship between the simplified acquisition threshold and policies to ban text messaging while driving?"

Complex questions require answers that must be expressed semantically; ideally in natural language.

Use Case 1. *Natural-Language Query: Simple Question*

This use case comprises the simple question types enumerated and described previously, posed in natural language by either a contracting officer or a representative of a prospective offeror. The goal of such a question might be to understand a document or requirement; determine the feasibility or legality of an action; or otherwise interact with acquisition regulations in a manner that can be expressed in formal logic. Examples:

User: Contracting Officer
Query: "What documents must be provided to sole-source a CPFF contract for services?"
—
User: Prospective Offeror
Query: "Is a business that is majority owned by a private equity firm eligible to subcontract on a Phase II STTR contract?"

Use Case 2. *Natural-Language Query: Complex Question*

This use case comprises the complex question types enumerated and described previously, posed in natural language by either a contracting officer or a representative of a prospective offeror. The goal of complex questions is primarily to understand a document or requirement, rather than to determine the feasibility or legality of an action. Unlike simple questions, complex questions are not well posed in terms of formal logic queries to an ontological database. Examples:

User: Contracting Officer
Query: "What is an example of an acceptable sole-source justification for a fixed-price subcontract issued by a prime under a cost-type services contract?"
—
User: Prospective Offeror
Query: "What happens during the contract negotiation phase?"

Use Case 3. *Requirement Specification and Flow-Down*

This use case comprises automatic specification from a solicitation of those compliance requirements that will become incumbent upon a successful offeror—whether for use by a contracting officer in developing a contract or by contractor personnel in assessing the risk/reward tradeoff of a potential opportunity. This specifically includes specification of possible acquisition strategies (viz., alternative authorities for acquisition, set-asides, sole-source, etc.)

Like Case 1, Case 3 expresses logical entailments, albeit based upon properties that are assessed automatically rather than in response to a question. Like Case 2, Case 3 requires (some) natural-language responses. While certain responses can be provided in terms of enumeration (e.g., a listing of required clauses), others require natural-language instruction/guidance.

Technical Approach

Architecture and Training

The use cases for MICA impose distinct requirements on the design and architecture with some areas of overlap. In particular, all use cases require some measure of natural-language processing (NLP) to parse, interpret, and machine-format queries or select and format natural-language output. However, our focus here is on those elements of MICA that enable those capabilities that are *particular* to it: acquisition-policy question-and-answer and specification of compliance requirements.

The technical requirements of Use Case 1, aside from NLP for ingestion of queries, can be addressed within a framework of formal semantics. In other words, the natural-language queries can be reframed as queries to an ontological database populated by ingesting the relevant source materials (viz., FAR, DFARS, etc.) and computing (first-order) logical entailments.

The technical requirements of Use Case 2 can be realized in at least two ways. First, a complex question can be understood as a "compound" or "composite" question that can be answered through (reasoning over and formatting) the answers to multiple simple queries. Alternately (*pace* Clark), a complex question can be understood as a linguistically formatted query that seeks to return a primary source document or information extracted from a primary source document—perhaps having undergone reformatting or combination through result ranking and/or formal reasoning.

Presently, the technical approach we are following in the development of MICA explicitly accounts for the distinction between these two use cases. The MICA architecture incorporates an encoding of acquisition rules, policy, and guidance in a structured knowledge base that supports formal reasoning and query (i.e., an ontology). It also follows the question-and-answer paradigm of IBM Watson, learning deep semantic knowledge from shallow syntactic information (viz., primary texts) and using this learned knowledge to interpret the semantic intent of questions and return relevant information extracted from source documents. Both are supported in extraction and encoding of information by domain-specific structured information (e.g., a legal ontology customized to address acquisition terminology).

Learning from Semi-Structured Content

Both components of MICA must ingest information from textual documents. But, unlike many conventional text documents, the acquisition rules, policies, and procedures are not unstructured prose. Legal text comprising the Code of Federal Regulations (CFR), FAR, DFARS, and other relevant regulations is semi-structured. Information is formatted in a hierarchical tree structure and heavily cross referenced between sections for the purpose of clarifying, supporting, interpreting, or limiting the requirements of the section in which the reference is found.

These references can be within or between documents and can be expressed in a structured manner, "see 48 CFR 9903.201-1," or in an ad-hoc manner that requires semantic interpretation, "using the rate specified in 26 U.S.C. 6621(a)(2)."

While this differs from the standard case, it supports extraction of semantics from syntax by giving a formal structure to the textual context.

Active Semi-Supervised Learning

Approaches such as Watson infer deep semantic information from shallow syntactic information contained in a corpus of documents through an unsupervised learning process (Ferrucci et al. 2010; McCord, Murdock, and Boguraev 2012; Fan, Kalyanpur, Gondek, and Ferrucci 2012) as do (some) approaches to build a knowledge base from text (Schubert 2002). However the question-and-answer capability for a given subject domain and corpus requires supervised training to achieve acceptable levels of performance (Ferrucci et al. 2010; Fan, Kalyanpur, Gondek, and Ferrucci 2012). This training is two-fold. First, it takes the conventional form of question-answer pairs (ideally drawn from the true query space) (Ferrucci et al. 2010; Fan, Kalyanpur, Gondek, and Ferrucci 2012). Second, it takes the form of fine-tuning the feature-extraction process—expert guidance of the algorithms through which the system extracts syntactic information and infers semantic information (Ferrucci et al. 2010; Fan, Kalyanpur, Gondek, and Ferrucci 2012). For example, a subject-matter expert (SME) might need to indicate which clauses of a contract or which portion of a Government Accountability Office (GAO) decision are more critical to parse or how a sentence structured in legal language should be parsed and interpreted. As Watson CTO and chief architect Sridhar Sudarsan puts it, "someone with expertise in the subject needs to identify from that corpus what bit is the right answer (Heath 2014)."

In our previous development of the SAGE semantic reasoning and suggestion engine, which integrated SME judgments with unsupervised machine learning to optimize the creation of simulation-based training scenarios, we found that the training process and the performance of SAGE could be enhanced by obtaining SME input within an active semisupervised framework. Semisupervised learning combines unsupervised learning of structures and correlations from raw, unlabeled data with supervised learning from selected, labeled data from SMEs. Active learning uses an algorithmic approach to determine which pieces of labeled data will be most informative and solicits those from SMEs. This both accelerates and improves the quality of the learning process.

For MICA we have developed an active-learning approach that operates by forming a dynamically updated ontology from the question-and-answer pairs input to and returned by the system. Through automated machine reasoning over this ontology to discover logical inconsistencies, SMEs are cued to provide targeted training that addresses failures in semantic inference. As Sudarsan has noted, training Watson "is not a one-time exercise, it's really an ongoing iterative approach (Heath 2014)." Our active semisupervised approach to SME engagement optimally supports that iterative learning. Similarly, the semisupervised architecture also supports integration of side information, such as user up/down voting, to improve response quality in a lifelong-learning approach.

Technical Challenges

Parsing Legal Language

The syntactic parsing (McCord, Murdock, and Boguraev 2012) structured databases (viz., DBpedia and WordNet) (Ferrucci et al. 2010) and semantic-inference algorithms systems like Watson rely upon (Fan, Kalyanpur, Gondek, and Ferrucci 2012) assume natural-language constructions and conventions for representing meaning. In contrast, legal documents, such as acquisition regulations are written in legal language that may provide a challenge to a Watson like system or conventional NLP parsing of the information into a structured semantic knowledge base (just as it does for many *people*).

Cross-Document Inference

Interpreting acquisition regulations requires cross-referencing between multiple documents: DFARS refers to the FAR, which in turn refers to USC. Moreover, regulations are subject to clarification (e.g., in memoranda from the Under Secretary of Defense) and interpretation (e.g., in United States GAO bid-protest decisions). While the design of Watson (and similar systems) allows for inference across multiple documents (Ferrucci et al. 2010; Fan, Kalyanpur, Gondek, and Ferrucci 2012), the ability to do so depends on (1) appropriate choice of the corpus contents such that they span the query space and (2) guidance from SMEs during feature-extraction learning. In MICA this is supported by a taxonomy of the query space within an ontological framework that supports automated selection of corpus documents based on SME input and the active semisupervised learning framework.

Inter- and Intradocument Consistency

Because of the size and complexity of acquisition regulations, inconsistencies can and do occur between documents and within documents as a result of imprecise language, modifications to language during the legislative process, conflicting agency interpretations of executive or intra-agency mandates, or simple error. Watson and similar systems assume the semantic information implicit in the unstructured documents of the corpus is consistent. But such systems typically do not check this information for consistency in the manner of a formal ontology (e.g., as in Wolfram Alpha), presuming that supervised training will provide sufficient information about relevant contexts to guide the algorithm to the appropriate data source given information implicit in the queries. Because MICA forms a formal ontology during training, it is able to direct attention to portions of the acquisition regulations that require such context-information distinction or may, in fact, be inconsistent. This is a significant side benefit of MICA. Currently, inconsistencies are often addressed through legal actions pursued through the GAO or the United States Court of Federal Claims. MICA allows inconsistencies to be addressed within agencies before protests arise.

References

Defense Science Board Task Force on Defense Industrial Structure and Transformation. 2008. Creating an Effective National Security Industrial Base for the 21st Century: An Action Plan to Address the Coming Crisis. Technical Report, Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics.

Heath, N. 2014. Five things you need to know before building an IBM Watson app. URL http://www.techrepublic.com/article/fivethingsyouneedtoknowbeforebuildinganibmwatsonapp/

Fan, J.; Kalyanpur, A.; Gondek, D. C.; and Ferrucci, D. A. 2012. Automatic knowledge extraction from documents. *IBM J. Res. & Dev.*, 56(3/4):5:1–10.

Ferrucci, D.; Jennifer Chu-Carroll, E. B.; Fan, J.; Gondek, D.; Kalyanpur, A. A.; Lally, A.; Murdock, J. W.; Nyberg, E.; Prager, J.; Schlaefer, N.; and Welty, C. 2010. Building Watson: An overview of the DeepQA project. *AI Magazine*, Fall:59–79.

McCord, M. C.; Murdock, J. W.; and Boguraev B. K. 2012. Deep parsing in Watson. *IBM J. Res. & Dev.*, 56(3/4):3:1–15.

Schubert, L. Can we derive general world knowledge from texts? M. Marcus (ed.), Proc. of 2nd Int. Conf. Hum. Lang. Tech. Res. (HLT 2002).

Acknowledgements

The work described here was performed as part of an effort sponsored by the Air Force Research Laboratory (AFRL).