Introspection and Adaptable Model Integration for Dialogue-based Question Answering

Daniel Sonntag
German Research Center for AI (DFKI)
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
sonntag@dfki.de

Abstract

Dialogue-based Question Answering (QA) is a highly complex task that brings together a QA system including various natural language processing components (i.e., components for question classification, information extraction, and retrieval) with dialogue systems for effective and natural communication. The dialogue-based access is difficult to establish when the QA system in use is complex and combines many different answer services with different quality and access characteristics. For example, some questions are processed by open-domain QA services with a broad coverage. Others should be processed by using a domain-specific instance ontology for more reliable answers. Different answer services may change their characteristics over time and the dialogue reaction models have to be updated according to that. To solve this problem, we developed introspective methods to integrate adaptable models of the answer services. We evaluated the impact of the learned models on the dialogue performance, i.e., whether the adaptable models can be used for a more convenient dialogue formulation process. We show significant effectiveness improvements in the resulting dialogues when using the machine learning (ML) models. Examples are provided in the context of the generation of system-initiative feedback to user questions and answers, as provided by heterogeneous information services.

1 Introduction

You are visiting a football match in Berlin and you take a mobile mini computer with you which is able to answer questions in real-time. If you ask, *"Who was world champion in 1990?"*, state-of-the-art question answering systems for this specific domain with a natural language understanding component and access to a knowledge base should be able to answer with great accuracy, *"That was Germany"*. Later, since you are new to the city, you are on a sightseeing tour. During the bus ride, you pass Castle Charlottenburg which arouses your curiosity, *"I wonder who might have built Castle Charlottenburg?"*

Unfortunately, most of the specific domain question answering systems would respond with *"No Answer"* after checking the knowledge base where the answer cannot be found (a task which might also consume a lot of time). In this situation, the user would be very dissatisfied with the system. Most existing approaches focus on improving the natural language understanding capability and/or the quality of the provided factual answers. Such improvements are important, but do not enhance the robustness of the system on a large scale. For example, in order to enhance the range of possible questions to be answered, open domain access using a search engine could be realised as a fallback strategy. This would potentially enhance recall, but also mean a loss of precision combined with problems of result provenance—the results are less reliable. Further problems occur if different information sources have different access characteristics, e.g., Web Services answering questions such as *"What's the weather like tomorrow?"* could be temporarily unavailable. This leads to a situation where efficiency, effectiveness, and naturalness of the question answering dialogue is hard to achieve.[1]

We focus on how to improve the QA system with a suitable dialogue within the QA dialogue and QA system's capabilities. In the following user-system dialogue example, adequate question feedback is shown in bold:

1. **U:** "When was Brazil world champion?"
2. **S:** "In the following 5 years: 1958 (in Sweden), 1962 (in Chile), 1970 (in Mexico), 1994 (in USA), 2002 (in Japan)." (**6000 ms**)
   Later ...
3. **U:** "What can I do in my spare time on Saturday?"
4. **S: "Sorry, services for such questions were unavailable a short while ago, shall I continue? This may take a minute or so ..."** (**600 ms**)
5. **U:** "Ah okay, I can wait."
6. **S: "Unfortunately, this service only produces empty results at the moment."** (**52000 ms**)

   Later on the bus ride ...

[1] Multi-strategy approaches use different QA subsystems when searching for answers to questions. An increasing number of open-domain QA systems have started using several retrieval approaches (e.g., by employing different search engines and different query expansions) and multiple answer extractors (e.g., keyword-based, concept-based, or based on user feedback, etc.). Particularly, the need for combining different data sources is of great importance.

7. **U:** "Now I would like to know who built Castle Charlottenburg."
8. **S: "It's not in my knowledge base. I will search the Internet for a suitable answer. Empty results are not expected, but the results won't be entirely certain." (600 ms)**
9. **S:** I think it was built by Johann Nehring and Martin Grünberg. (**10000 ms**)

   Later, in front of the stadium ...
10. **U:** The French team is just getting off the bus. "Can I have pictures and names of all French players?"
11. **S: "I predict fifty seconds for the search and download of the image content. The knowledge base will be queried immediately." (600 ms)**

In order to model this (meta) dialogue[2], the system must be able to predict empty results, answer times, and classify queries for the probability of success according to query features and specific access/quality properties of the answer services in a changing environment. What we aim at with our work is to provide such a dialogue-based question answering functionality by employing an introspective mechanism based on ML for the generation of adaptable reasoning models. These allow the dialogue manager to monitor and control itself. Subsequently, we will show methods for evaluating the new methodology by improving the dialogical feedback in dialogue-based question answering.

2 Related Work

Many research projects aim to enhance the user's satisfaction in dialogue systems by developing new forms of adaptivity management which complement multimodality and multilinguality. Recent research in dialogue systems focuses on adaptable dialogue management strategies. According to [Walker, 2000] and [Levin *et al.*, 2000], dialogue strategies similar to those designed by human experts can be learned in the Markov Decision paradigm with reinforcement learning. (This was used on the dialogue task level; it showed that large state spaces with more than about five non-binary features a hard to deal with.) Further advances have been made by natural multimodal dialogue systems (see [van Kuppevelt *et al.*, 2007]), and by hierarchical reinforcement learning and dialogue simulations toward adaptable dialogue management strategies. Probabilistic methods in spoken dialogue systems, e.g., [Young, 2000], emphasise the importance of feature aggregation and filtering in order to obtain sufficiently small state spaces while still conveying the decision-relevant information. Dialogue simulations have been proposed to obtain enough training data. [Mollá and Vicedo, 2007] provide a list of additional question answering systems in restricted domains; [Basili *et al.*, 2007] propose a system with the outstanding feature of robustness through adaptive models of speech recognition and planning of dialogue moves; [Maybury, 2003] proposes a roadmap for question answering, dealing with resources to develop or evaluate question answering, as well as methods and algorithms. Interactive/dialogue-based, multimodal, and constrained question answering (in terms of resources and solutions) are among the longer term objectives.

In the context of resource-bounded reasoning (i.e., to embed complex reasoning components in real-world applications), it is especially the computational commodities, such as time requirements, which resemble the requirements of our information gathering QA application (e.g., see [Zilberstein and Russell, 1992]). Because of the increased level of deliberation in speech-based communication for the QA process, techniques including anytime algorithms or time-bounded search could be addressed. However, this requires the introspective mechanism to have access to the whole message scheduling process of the QA submodules for initiating a request. Our adaptable model is constrained in the way that it remains on the dialogue manager side. In the BBN system by [Mulvehill *et al.*, 2007], an adaptation module provides and refines models to account for changes in the world state and to improve the execution of plans by mapping failure symptoms to causal faults, such as incorrect model parameters. In addition, to allow the task system to detect and recommend models, we try to automatically repair models, but have the advantage that false positives cannot affect the task level (QA process), remaining instead on the dialogue level.

Although some question answering systems exist[3] which employ dialogue with advanced technical approaches (i.e., empirical, linguistic, and knowledge based), methodologies dealing with increasing system complexity and changing resource availability have yet to be developed. Since the above-mentioned existing dialogue models are not well-suited for special meta level dialogue phenomena in question answering applications (for example, providing processing feedback as in the example dialogue), alternative dialogue models are needed.

A new adaptation model (section 4) is implemented by a reaction and presentation module (REAPR) which manages the dialogical interaction for the supported dialogue phenomena, such as flexible turn-taking, incremental processing, and the adaptation of the action rules. Our approach differs from other information state (IS) approaches (e.g., [Matheson *et al.*, 2000]) by generating information state features from the ontological instances generated during dialogue processing. Ontological structures that also may change over time vastly enhance the representation capabilities of dialogue management structures.

3 QA Architecture

Figure 1 shows our QA architecture and the information servers. We face a QA Information Integration Problem in the latter. Information integration for deriving answers must be done while considering multiple heterogeneous multimedia repositories dealing with structured, semistructured, and unstructured resources (O, T, A, O). Heterogeneous data sources have different access, reliability, and trust characteristics. Especially those involving different data quality characteristics of heterogeneous data sources demand data/information metamodels. For example, the open domain QA engine (e.g., as a fallback strategy) enhances recall, but

[2]The dialogue provided derives from application potentials of Dialogue and Semantic Web Technology. An interactive demo is provided by Deutsche Telekom and is available at http://smartweb.dfki.de/SmartWeb_FlashDemo_eng_v09.exe.

[3]For example, see the Halo (http://www.projecthalo.com/), SmartWeb (http://www.smartweb-projekt.de), and BirdQuest [2004] question answering systems.

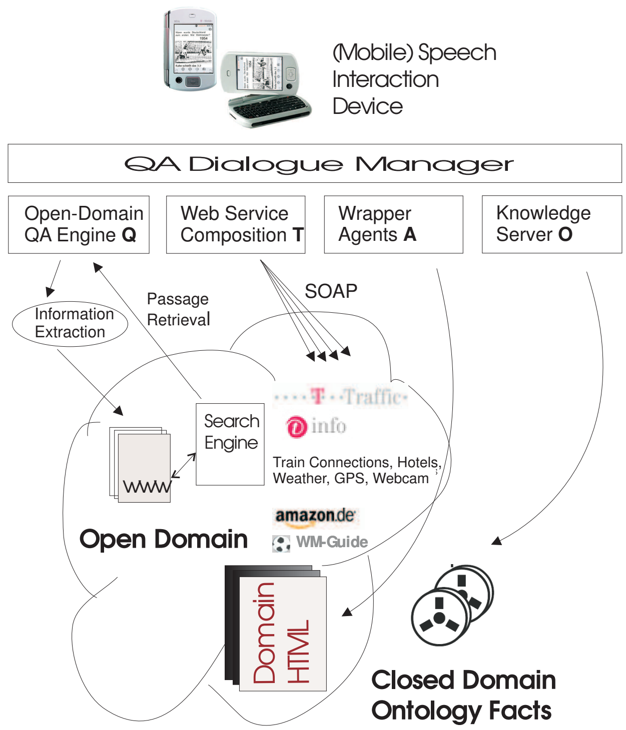

Figure 1: The Basic QA Architecture and Information Servers

also means a loss of precision combined with problems of metadata representation for result provenance. In our QA system [Sonntag *et al.*, 2007], the responsibility for meaningful metadata, such as confidence values, lies with the delivering components (open-domain QA, WS composition, HTML wrappers, and a knowledge server). All information servers have been appealed for delivering confidences and an explanation from what these confidences developed. Metadata can be mined at the dialogue modeller's option for reaction and presentation decisions. With the help of the generated models, the system can be tuned to detect and communicate uncertainties in the QA results on the dialogue level when using the (mobile) speech-based interaction device.

4 Dialogue Manager Model

The dialogue manager model is based on ontological knowledge representation structures, as introduced for semantic-based applications by [Fensel *et al.*, 2003]. These structures make up the assertions of the current dialogue turn and the dialogue history, i.e., the semantic query as partly-filled ontology instances; the results of the different answer services; answer status information such as the elapsed time and result confidence; and the generated answer to be presented on the interaction device. All these information state features follow the structural commitments of ontology-based representation suggestions on foundational and domain models, as described in [Oberle *et al.*, 2007]. The ontological features for information states as shown in table 1 are observed and extracted as input features for learning dialogue reaction decisions.

Feature Class	IS State Features
ASR	*Listening, Recording, Barge-in, Last-ok,*
NLU	*Confidence, Domain relevance*
Query	*Dialogue act, Question foci, Complexity, Context object, Query text*
Answer	*Success, Speed, Answer streams, Status, Answer type, Content, Answer text*
Manager	*Turn/Task numbers, Idle states, Waiting for results, User/system turn, Elapsed times: input/output, Dialogue act history (system and user) e.g. reject, accept, clarify*

Table 1: IS Feature Classes and Features: Automatic Speech Recognition (ASR), Natural Language Understanding (NLU); the query, answer, and dialogue manager features.

The self-understanding of what the system can or cannot do at any moment is crucial and includes predicting the confidence of results and the availability of information services. Error handling (preferably before errors occur) through *thinking ahead* can be seen as a key factor to increasing general acceptance, usability, and naturalness of the dialogue-based interface. This is what the adaptation model should provide for us and it is obtained by running dialogue sessions.[4] The adaptation model implemented by the reaction and presentatation model is shown in figure 2.

Information states are traditionally divided into global and local variables which make up the knowledge state at a given point of time. We also use this global and local representation to differentiate between global dialogue information about the dialogue session (such as user context information) and local information about the current user or system turn, e.g., question type information. Thereby, the observed global and local ontological assertion instances (abox) are hosted in the dialogue information hub (iHUB). The ontology instances, represented by the IS state features, are translated into propositions of propositional calculus using an indicator function $I_B(feature)$ on the local and global record sets B_{local} and B_{global}. For example, $DM\_QUERY\_FOCUS = sportevent\#Match$ means that the current query processed in the dialogue manager contains a sportsevent match instance as question focus.

The next step is very important. The information state delivers contemporary feature-based information about the ongoing dialogue and internal processes. By applying special triggers and using database schemata, the extraction of data sets can be controlled to a great extent. *Trigger events* are special bindings of variables in the dialogue state we observe. This means we observe the abox (after every lock-for-write command) for occurrences of specific ontological instances which fire a rule to produce a new learning instance.

[4]Unlike traditional Wizard-of-Oz experiments where the expert needs to manually build appropriate modelling rules, we will try to automate the model creation and integration process. The predictive models should provide a kind of think-ahead functionality to obey dialogue reaction and presentation constraints.

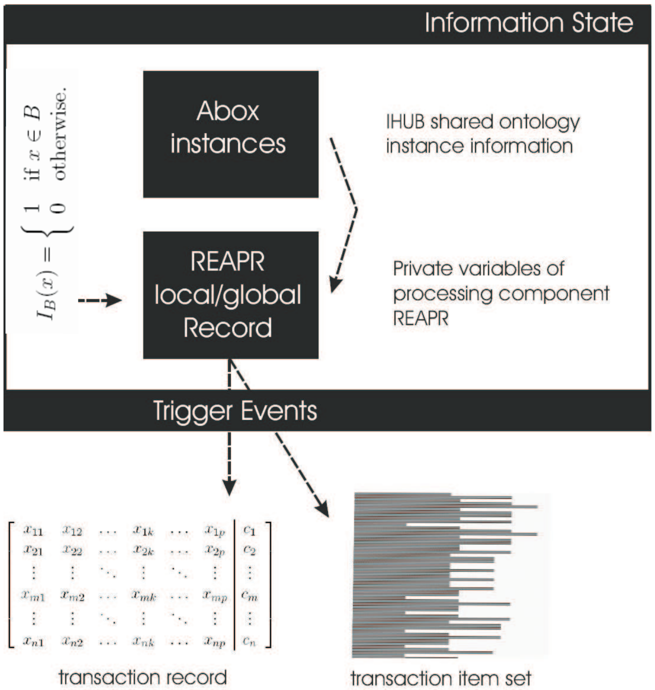

Figure 2: Dialogue Manager and Adaptation Model

With the help of trigger events, we extract a feature vector containing the dialogue IS state features at the time when the trigger fires. The rationale for selecting these features is the expectation that some of these features properly describe a certain dialogue situation according to the feature values (which correspond to parts of the dialogue information state). We used different trigger types to extract transaction records for supervised classification, or transaction item sets for association analysis. Time triggers extract a new record or item set as the dialogue proceeds every 500ms, for example. Event triggers use certain events, e.g., the completion of the semantic query, or specific result structures from the information services, to fire. After a series of different experiments where we compared the predictive power of 50 data sets with different time and event triggers (we used attribute entropy, as well as jointed and conditional entropy measures, information gain, and the Gini Index) we determined that event triggers worked best when the receipt of a new single result obtained from any information service was taken as the trigger event.

In this way, optimisation problems can be formulated for very specific decisions in dialogue management (due to a basic finite state automaton providing the basic QA dialogue control); datasets can also be collected. In addition, the methodology we use (actually, we adhere strictly to the knowledge discovery in databases (KDD) process) includes effective preprocessing of feature relevance. In addition, the cost complexity of, e.g., the employed K-Nearest Neighbour and Naïve Bayes classifiers, are bound to $O(nm)$. Generally, medium-sized training sets, where we collect less than 1000 instances, obtained the most useful predictive ML models. This will be evaluated next in this paper.

5 Evaluation and Model Integration

We would like to introduce a statement that might be read as the overall hypothesis:

Simple Machine Learning models could be employed on the meta-level to reason about the environment input of an (conversational) agent and adapt the dialogue.

Through the evaluation, we wanted to gather two results. First, we wanted to see whether the learned models have accurate performance in terms of the classification performance measures (i.e., accuracy and Area-Under-Curve for ROC curves). Second, we wanted to verify that dynamically generated models can be effectively used for improving question feedback, predicting answer times, or presenting reliable open-domain QA results. Otherwise, users may be frustrated due to discourse constraints the system is not able to provide. For example, the level of dialogue initiative is one of the most important reaction constraints for information seeking dialogue systems. Key aspects of system initiative include:

- maintaining the dialogue with the user by reporting on the question processing status.
- informing the user about the probability of query success, i.e., the probability the user will be presented the desired information.
- informing the user as to why the current answering process is due to fail.

According to this, dialogue integration basically means detecting and communicating uncertainties in the results. Our training examples are collected from real user interactions with our baseline dialogue system (also explained in the summative evaluation of the SmartWeb system [Mögele and Schiel, 2007]). We conducted experiments on transaction items sets and transaction records. Here, we focused on experiments with transaction records, which have a supervised or self-supervised target variable (according to the ontology assertions). Self-supervision is to be understood as the ability to bi-directionally interact with the environment (by introspection), and to include exploratory (ontology-based) metadata into the internal decision process (in our case, dialogue management duties). This means the dialogue engineer selects an interesting target variable which some dialogue rules are based on.[5]

To illustrate, we have selected 2 different question sets which are run one after the other. In the first set, we have 250 user requests about the football application domain. Second, we have 88 user requests of the open domain, together with a supervised target which states which of the four answer streams O, T, A, Q would be most suitable for providing the answer. (All questions are similar to the example in the

[5] In other experiments, we used the transaction item sets. Association rule learning is a typical data mining technique. Association rules are expressions like $X \Rightarrow Y$, where X and Y are disjoint sets of items. We studied the predictive ability of association rule measures we use in our experiments such as *support*, *confidence*, *lift*, and *conviction*. Association rules can then be used to induce rules for incremental result presentation, to enhance recall for semantic questions (both similar to the presented classification models), and to direct an incremental learning and adaptation strategy.

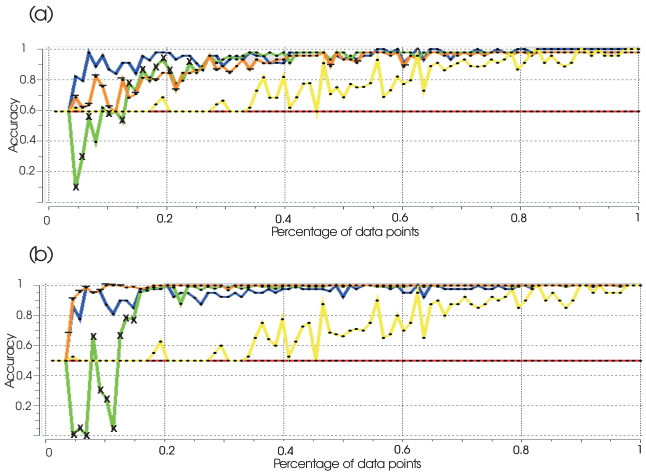

Figure 3: Learning curves on a supervised data set: Naïve Bayes (Red -), SVM (Green x), k-NN (Blue *), C4.5 (Yellow *); the *x*-axis shows the percentage of data points used for learning, the *y*-axis shows the predictive performance (*accuracy*) at 10-fold cross-validation; (a) shows the learning of the positive model (result is non-empty); (b) shows the learning of the negative model (result is empty).

introduction.) We have trained four classifiers in these four classification targets.

5.1 Supervised Classification Models

In these experiments, we used the generated transaction records. All results presented here are statistically significant over the baseline majority vote (according to ROC) based on 10-fold cross validation and 10-fold bootstrap estimates for *all* classifiers. While drawing ROC curves, we compared their predictive performance. In addition, confidence intervals were computed at $p = 0.05$.

Predict Answer Times

We were interested in the learning curves for this classification problem. Learning curves reveal how many instances are needed for a considerable predictive accuracy on this as well as on similar classification problems. Especially in user interactive, online environments, learning curves are good indicators of the time needed to execute an adaptation process. In addition, learning curves indicate which algorithms are best suited for a particular learning problem. Figure 3 shows the learning rates on the supervised data set.

Only a few user questions (e.g., real users questioning the dialogue system) are needed to learn the non-accessibility concept; using only 20% of the second data set allows us to predict a non-empty answer with an accuracy of above 90% (see figure 3 (a)).

---

1. **U:** "What can I do in my spare time on Saturday?"
2. **S: "Sorry, these types of questions cannot be answered at the moment ..."** (**500 ms**)
3. **U:** "Oh I see, ... "

Provide Question Feedback

During the formative evaluation, we recorded the following dialogue: "Can you answer questions? I am on this bus ride in Berlin and only want to know who built Castle Charlottenburg and after a minute you say 'Charlottenburg Palace? Damn!"

In order to improve the question feedback, the following action models can be learned while using the first data set:

1. **"It is not in my knowledge base"** (stream/answer time prediction);
2. **"I will search the Internet for a suitable answer"** (stream prediction);
3. **"Empty results are not expected, but"** (self-supervised answer prediction);
4. **"... the results won't be entirely certain"** (NLU model + expert Q model).

In (3), self-supervised answer prediction means that we were able to distinguish between empty answers (e.g., "zero goals") according to the closed-world assumption of the football domain and non-successful answering processes by exploiting the ontological answer structures. In (4), the NLU model is combined with an expert model on Q answers; roughly speaking, open-domain answers are not reliable.

Learn to Present Reliable Open-Domain Results

We were after a possibility of how to differentiate between a question the open-domain QA engine can reliably answer and which it cannot. The corresponding classification task and learned model revealed that the semantic answer type "Person" is predicted to be highly confident for Q questions using our open-domain QA engine. As a result, we present and synthesise the answer snippet directly, instead of presenting a top-ranked list. In speech-based QA systems, the result is very impressive:

---

1. **U:** "Who is the German Chancellor?"
2. **S:** "Who is the German Chancellor?" (**900 ms**)
3. **S:** "I will search the Internet for a suitable answer." (**900 ms**)
4. **S: "Angela Merkel."** (+ display of a short supportive text in the media screen) (**12000 ms**).

5.2 Verification

On six point Likert scales, we verified the approach before and after the adaptation. We did not conduct a second formative evaluation. Instead, we presented the adapted dialogues to 10 human judges. Since our false-positive-rate for the initiation of a question or answer feedback is extremely low, the adapted system received a higher overall score due to much better ratings on the user evaluation questions: "The error messages are helpful."; (b) "The question processing is fast enough."; (c) "The system leads the user quickly to the desired information.", (d) "The pauses between question input and answer output seem to be short.". In their surveys, the users reported that especially the long response times for some queries (if longer than 15 sec.) are perceived much shorter when the question feedback is adequate. The difference is statistically significant at $alpha = 0.05$ using an

ANOVA test, indicating the results as positive outliers compared to the formative evaluation of the base system. This verifies that the adaptation process is useful.

6 Conclusion

In our work, we implemented an approach to introspect a dialogue-based QA system. The learned models can be integrated into the dialogue manager decision process for automatically providing feedback on questions and answers. Overall, the positive effect of question and answer feedback can be easily seen in the evaluation examples. In addition, by using the meta dialogue principle, we are only adding additional information thereby maintaining the level of QA competence in the quality of the answers. The automatic selection of the appropriate models, however, remains a challenge. As shown in the last example (Learn to Present Reliable Open-Domain Results) we are also able to direct the presentation of QA results and highlight reliable open-domain results. This is very effective in speech-based systems. It provides a solution to one major problem in QA systems: adequate question and result feedback. In this respect, the application of machine learning at the meta level showed significant and understandable improvements to the overall QA task.

Since our approach can be used with any unsupervised question set (and supervised with one target variable) due to the ontological features in the query and result structures, we can easily apply it to other question sets to generate new introspective rules.

Acknowledgments

This research has been supported by the THESEUS Programme funded by the German Federal Ministry of Economics and Technology (01MQ07016). I would also like to thank the SmartWeb partners, as well as Wolfgang Wahlster and Randy Goebel for useful comments.

References

[Basili *et al.*, 2007] Roberto Basili, Diego De Cao, and Cristina Giannone. *On the Move to Meaningful Internet Systems: OTM 2007 Workshops*, chapter Ontological Modeling for Interactive Question Answering, pages 544–553. Springer Berlin / Heidelberg, 2007.

[Fensel *et al.*, 2003] Dieter Fensel, James A. Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. MIT Press, 2003.

[Jönsson *et al.*, 2004] Arne Jönsson, Frida Andén, Lars Degerstedt, Annika Flycht-Eriksson, Magnus Merkel, and Sara Norberg. Experiences from combining dialogue system development with information extraction techniques. In Mark T. Maybury, editor, *New Directions in Question Answering*, pages 153–168. AAAI Press, 2004.

[Levin *et al.*, 2000] E. Levin, R. Pieraccini, and W. Eckert. A stochastic model of human-machine interaction for learning dialog strategies. In *IEEE Trans. on Speech and Audio Processing, Vol. 8*, 2000.

[Matheson *et al.*, 2000] C. Matheson, M. Poesio, and D. Traum. Modelling grounding and discourse obligations using update rules. In *Proceedings of NAACL 2000*, 2000.

[Maybury, 2003] Mark T. Maybury, editor. *New Directions in Question Answering, Papers from 2003 AAAI Spring Symposium, Stanford University, Stanford, CA, USA*. AAAI Press, 2003.

[Mögele and Schiel, 2007] Hannes Mögele and Florian Schiel. Summative Evaluation of the SmartWeb Prototype 1.0. Technical report, Ludwig-Maximilians-Universität Munich, September 2007.

[Mollá and Vicedo, 2007] Diego Mollá and José Luis Vicedo. Question Answering in Restricted Domains: An Overview. *Comput. Linguist.*, 33(1):41–61, 2007.

[Mulvehill *et al.*, 2007] Alice M. Mulvehill, Brett Benyo, Michael T. Cox, and Renu Kurien Bostwick. Expectation failure as a basis for agent-based model diagnosis and mixed initiative model adaptation during anomalous plan execution. In Manuela M. Veloso, editor, *IJCAI*, pages 489–494, 2007.

[Oberle *et al.*, 2007] Daniel Oberle, Anupriya Ankolekar, Pascal Hitzler, Philipp Cimiano, Michael Sintek, Malte Kiesel, Babak Mougouie, Stephan Baumann, Shankar Vembu, Massimo Romanelli, Paul Buitelaar, Ralf Engel, Daniel Sonntag, Norbert Reithinger, Berenike Loos, Hans-Peter Zorn, Vanessa Micelli, Robert Porzel, Christian Schmidt, Moritz Weiten, Felix Burkhardt, and Jianshen Zhou. DOLCE ergo SUMO: On foundational and domain models in the SmartWeb Integrated Ontology (SWIntO). *Web Semant.*, 5(3):156–174, 2007.

[Sonntag *et al.*, 2007] Daniel Sonntag, Ralf Engel, Gerd Herzog, Alexander Pfalzgraf, Norbert Pfleger, Massimo Romanelli, and Norbert Reithinger. *SmartWeb Handheld—Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services*, volume 4451 of *Lecture Notes in Computer Science*, pages 272–295. Springer, 2007.

[van Kuppevelt *et al.*, 2007] Jan van Kuppevelt, Laila Dybkjaer, and Niels Ole Bernsen. *Advances in Natural Multimodal Dialogue Systems (Text, Speech and Language Technology)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[Walker, 2000] Marilyn A. Walker. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416, 2000.

[Young, 2000] Steve Young. Probabilistic methods in spoken dialogue systems. *Philosophical Transactions of the Royal Society (Series A)*, 358(1769):1389–1402, 2000.

[Zilberstein and Russell, 1992] Shlomo Zilberstein and Stuart J. Russell. Efficient resource-bounded reasoning in AT-RALPH. In *Proceedings of the First International Conference on AI Planning Systems*, pages 260–266, 1992.