

Abstract

Acquiring domain-specific knowledge necessary for creating a dialog system in a new task-oriented domain is a time consuming task that requires domain expertise. This dissertation explores the feasibility of using a machine learning approach to infer the required domain-specific information automatically from in-domain conversations. In order to achieve this goal, two problems need to be addressed: 1) creating a dialog representation that is suitable for representing the required domain-specific information, and 2) developing a machine learning approach that uses this representation to capture information from a corpus of in-domain conversations.

In order to solve the first problem, I propose a form-based dialog structure representation incorporating a three-level structure of *task*, *sub-task*, and *concept*. These components are observable in human dialogs. In terms of representation, tasks and sub-tasks are represented by forms while concepts are slots in a form. The notion of *form* is generalized as a repository of related pieces of information so that it can be applied to various types of task-oriented domains. Dialog structure analysis and an annotation experiment are used to demonstrate that the form-based representation has all the required properties: *sufficiency*, *generality*, and *learnability*. The proposed representation is applied to six disparate task-oriented domains (air travel planning, bus schedule inquiry, map reading, UAV flight simulation, meeting, and tutoring). While the form-based approach shows some limitations, it is sufficient to model important phenomena in dissimilar types of task-oriented dialogs, and thus has both *sufficiency* and *generality*. The annotation experiment shows that the form-based dialog structure representation can be applied reliably by novice annotators which implies that the representation is unambiguous and *learnable*.

For the second problem, inferring the form-based dialog structure representation from a corpus of in-domain conversations, I divide this dialog structure acquisition problem into two sub-problems, concept identification and form identification, to make the problem tractable. In order to identify a set of domain concepts, two unsupervised concept clustering approaches are investigated: statistical-based clustering and knowledge-based clustering. For most statistical-based clustering algorithms, we are able to find automatic stopping criteria that yield close to optimal results. The statistical-based approaches, which utilize word co-occurrence statistics such as mutual information and the Kullback-Liebler distance, while able to capture domain-specific usage of concept words cannot accurately identify infrequent concept words due to a sparse data problem. On the other hand, the knowledge-based approach, which uses semantic information stored in the WordNet lexical database, can identify domain concepts very accurately, but on the condition that the concepts are present in the knowledge source.

To determine different types of forms and their associated slots, a dialog is first segmented into a sequence of sub-tasks by an unsupervised text segmentation algorithm. Then, the bisecting K -mean sub-task clustering algorithm is used to group the sub-tasks that represent the same form type into the same cluster. Finally, a set of slots that is associated with each form is determined from the concepts present in each cluster. To handle fine-grained segments in spoken dialogs, TextTiling and HMM-based segmentation algorithms are augmented with a data-driven stop word list and distance weights. With these modifications significant improvement is achieved.

Even though the performance of the bisecting K -mean sub-task clustering algorithm can be affected by inaccurate sub-task boundaries, I found that moderate segmentation accuracy is sufficient for identifying frequent form types. Similarly, moderate sub-task clustering accuracy is sufficient for identifying essential slots in each form.

The results of both dialog structure acquisition problems, concept identification and form identification, show that it is feasible to acquire the domain-specific knowledge necessary for creating a task-oriented dialog system automatically from a corpus of in-domain conversations using unsupervised learning approaches. This data-driven approach could potentially reduce human effort in developing a new task-oriented dialog system.