

FLORIDA INTERNATIONAL UNIVERSITY
Miami, Florida

STATISTICAL DIALOG MANAGEMENT FOR HEALTH INTERVENTIONS

A dissertation submitted in partial fulfillment of the
requirements for the degree of
DOCTOR OF PHILOSOPHY
in
COMPUTER SCIENCE
by
Ugan Yasavur

2014

To: Dean Amir Mirmiran
College of Engineering and Computing

This dissertation, written by Ugan Yasavur, and entitled Statistical Dialog Management for Health Interventions, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Naphtali Rische

Tao Li

Mark Williams

Ubbo Visser

Christine Lisetti, Major Professor

Date of Defense: July 09, 2014

The dissertation of Ugan Yasavur is approved.

Dean Amir Mirmiran
College of Engineering and Computing

Dean Lakshmi N. Reddi
University Graduate School

Florida International University, 2014

© Copyright 2014 by Ugan Yasavur

All rights reserved.

DEDICATION

To my parents.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Christine Lisetti for her great support and advise during all these years. Thank you for always being available, constructive and for providing great advice.

I also want to thank the other members of my Ph.D Committee, Dr. Naphtali Rische, Dr. Tao Li, Dr. Ubbo Visser, and Dr. Mark Williams for their time taken in reading this dissertation, for their support and valuable feedback.

Finally, I want to thank all members of the Affective Social Computing Laboratory (ASCL). ASCL was like my home during all these years and I made great friends along the way.

ABSTRACT OF THE DISSERTATION
STATISTICAL DIALOG MANAGEMENT FOR HEALTH INTERVENTIONS

by

Ugan Yasavur

Florida International University, 2014

Miami, Florida

Professor Christine Lisetti, Major Professor

Research endeavors on spoken dialogue systems in the 1990s and 2000s have led to the deployment of commercial spoken dialogue systems (SDS) in microdomains such as customer service automation, reservation/booking and question answering systems. Recent research in SDS has been focused on the development of applications in different domains (e.g. virtual counseling, personal coaches, social companions) which requires more sophistication than the previous generation of commercial SDS. The focus of this research project is the delivery of behavior change interventions based on the brief intervention counseling style via spoken dialogue systems.

Brief interventions (BI) are evidence-based, short, well structured, one-on-one counseling sessions. Many challenges are involved in delivering BIs to people in need, such as finding the time to administer them in busy doctors' offices, obtaining the extra training that helps staff become comfortable providing these interventions, and managing the cost of delivering the interventions. Fortunately, recent developments in spoken dialogue systems make the development of systems that can deliver brief interventions possible.

The overall objective of this research is to develop a data-driven, adaptable dialogue system for brief interventions for problematic drinking behavior, based on reinforcement learning methods. The implications of this research project includes, but are not limited to, assessing the feasibility of delivering structured brief health

interventions with a data-driven spoken dialogue system. Furthermore, while the experimental system focuses on harmful alcohol drinking as a target behavior in this project, the produced knowledge and experience may also lead to implementation of similarly structured health interventions and assessments other than the alcohol domain (e.g. obesity, drug use, lack of exercise), using statistical machine learning approaches.

In addition to designing a dialog system, the semantic and emotional meanings of user utterances have high impact on interaction. To perform domain specific reasoning and recognize concepts in user utterances, a named-entity recognizer and an ontology are designed and evaluated. To understand affective information conveyed through text, lexicons and sentiment analysis module are developed and tested.

TABLE OF CONTENTS

CHAPTER	PAGE
1. Introduction	1
1.1 Statement of the Problem	3
1.2 Specific Research Questions and Objectives	4
1.3 Outline of the Dissertation	6
2. Literature Review	8
2.1 Dialogue Management Overview	9
2.1.1 Dialogue Management based on Machine Learning	11
2.1.2 Dialogue Management and Reinforcement Learning	13
2.2 Dialogue Management in Health Dialogue Systems	16
2.3 Spoken Dialog Systems	18
2.4 Embodied Conversational Agents and Spoken Dialog Systems	21
2.5 Brief Interventions for Alcohol Problems	24
3. Spoken Dialogue System Based on Markov Decision Processes	36
3.1 Approach	36
3.2 Reinforcement Learning Background	38
3.3 Markov Decision Processes	39
3.4 Modeling World with Interconnected MDPs	45
3.5 Agent and Dialogue Strategy Learning	46
3.6 Reward Function Design	50
3.7 Speech Recognition and Language Model	50
3.8 Dialog Corpus	53
4. Spoken Dialogue System Based on POMDP and Health Screening Dialogues	54
4.1 Underage Drinking Problems and Computer-based Interventions	55
4.2 Alcohol Screening And Brief Intervention For Youth	56
4.3 System Architecture Overview	57
4.4 Approach	58
4.4.1 Belief Monitoring and Partially Observable Markov Decision Processes	60
4.4.2 Representing Patterns in Health Interviews As Programatic Objects . .	62
4.5 Discussion	66
5. Experiments with Optimized and Random Dialog Policies	67
5.1 Sample Population	67
5.2 Evaluation Results	69
5.2.1 Task completion evaluation	69
5.2.2 Dialogue Evaluation	72
5.2.3 Subjective Evaluation	73
5.3 Implications	77
5.4 Discussion	79

6. Recognizing Named-Entities and Domain Specific Reasoning For Behavioral Health	81
6.1 Background in Named Entity Recognition	82
6.2 Approach	84
6.3 Ontology Design	85
6.4 WordNet	87
6.5 Architecture	89
6.6 Data and Evaluation	90
6.7 Discussion	93
7. Lexical Resources & Semantic Networks	95
7.1 Lexical Resources & Semantic Networks	95
7.2 Introduction	96
7.3 Background on Lexical Resources	99
7.4 Storing Semantic Data As A Graph	101
7.5 Core Structure of Unified Semantic Network	103
7.6 Affective Knowledge	106
7.7 Creating New Lexicons for Prior Sentiment Polarities & Affective Ratings	108
7.8 Evaluation	112
7.8.1 Prior Sentiment Polarities	112
7.8.2 Affective Ratings	113
7.9 Discussion and Implications of Unified Semantic Network	114
8. Sentiment Analysis	115
8.1 Introduction	116
8.2 Related Work	118
8.3 Approach	120
8.3.1 Named-Entities and Topic Modeling	120
8.3.2 Features and Interactions Between Words	122
8.4 Experiments	124
8.4.1 Data	125
8.4.2 Compared Methods	125
8.4.3 Experiment Results	127
8.5 Discussion	128
9. Conclusions	129
9.1 Summary	129
9.2 Future Directions	130
Bibliography	133
A. List of Dialog Actions of Spoken Dialog System	157
VITA	169

LIST OF FIGURES

FIGURE	PAGE
2.1 Dialogue System Architecture	9
2.2 Reinforcement Learning with MDP	14
2.3 Multimodal Embodied Conversational Agent Interface	23
2.4 First Step of Alcohol Screening and Brief Intervention	26
2.5 Second Step of Alcohol Screening and Brief Intervention	27
2.6 Third Step of Alcohol Screening and Brief Intervention for At-Risk Drinking	28
2.7 Third Step of Alcohol Screening and Brief Intervention for Alcohol Use Disorders	29
3.1 Representation Of World Model With MDPs	45
3.2 System Architecture and Dialogue Manager	47
4.1 Spoken Dialogue System Architecture for POMDP	58
4.2 Sample Dialogue Structure Designed With Dialogue Tool	64
5.1 Growth of Q-Values for Each Policy	73
5.2 Subjective Evaluation	75
5.3 Assessment - Negative (1) Neutral (4) Positive (7)	76
6.1 Behavioral Health Ontology	85
6.2 The System Architecture	89
7.1 Sample of Structure of the Graph	104
7.2 A Snapshot from Query Result Visualizer	105
7.3 An Example from the Graph	109
8.1 System Architecture	119
8.2 Dependencies	122

LIST OF TABLES

TABLE	PAGE
2.1 Sample Dialogue	31
3.1 State attributes and values for Step 1: Asking About Alcohol Use	40
3.2 Dialog Actions For Step 1 Question1	41
3.3 Explaratory Policies For Step 1 Question 1	44
5.1 Task Completion Rate: Training versus Testing	70
5.2 Subjective Evaluation Categories: Training versus Testing	77
6.1 Behavioral Named Entity Recognizer Evaluation Results	92
8.1 Node Features	123
8.2 Dependency Features	124
8.3 Experiment Results for Sentiment Analysis	127
A.1 Step 1 Dialog Actions	157
A.2 Step 2 Abuse Dialog Actions	160
A.3 Step 2 Dependence Dialog Actions	163
A.4 Step 3 At-Risk	167

CHAPTER 1

Introduction

Substantial amount of research has been conducted on developing dialogue systems in the last two decades, and as a result, commercial systems are being deployed in microdomains such as reservation/booking, customer center automation, or question answering systems. More recently, researchers have been concentrating on the development of dialogue systems which are able to take initiative, reason and infer while conducting dialogues in microdomains. These latest developments have lead to a high interest in developing dialogue systems for real world applications such as personal coaches, companions and virtual counselors, that are more sophisticated than reservation chatbots. The focus of this research project is the delivery of behavior change interventions based on the brief intervention counseling style via spoken dialogue systems. Although our approach is adaptable to other type of lifestyle issues (e.g. over-eating, lack of exercise, drug use), in this project we focus on excessive alcohol consumption as a target behavior.

Excessive alcohol consumption is regarded as a very worrisome public health problem in the USA: with approximately 85,000 of directly or indirectly attributable deaths per year, excessive alcohol use is the 3rd leading lifestyle-related cause of death in the United States [AJDJ04]. In 2006, there were more than 1.2 million emergency room visits and 2.7 million physician office visits due to excessive drinking [BHS⁺11]. Excessive alcohol use is also a risk factor for many health and social problems, including motor-vehicle crashes, violence, suicide, hypertension, unsafe sex, or unintended pregnancy. The economic costs of excessive alcohol consumption in 2006 were approximately \$223.5 billion [BHS⁺11]. To attempt to address these alarming statistics, health-care research has led to the development and deployment of behavior change interventions that can be delivered efficiently in primary care offices.

Brief interventions (BI) are evidence-based, short, well structured, one-on-one counseling sessions, focuses on specific aspects of target behavior, and are ideally suited for people with unhealthy lifestyles that they would need to change to remain healthy, for example drinking in ways that are harmful or abusive [NIA06].

BIs can be delivered in 3-5 minutes [MFSV02] and (for alcohol consumption as a target) aim to moderate a person's alcohol consumption to reasonable levels and to eliminate harmful drinking behaviors. BIs are the top ranked out of 87 treatment styles in terms of efficiency [MW02]. It is reported that even a few minutes of advice and discussion about behavioral problems can be as effective as more extended counseling [BG92]. Many challenges are involved in delivering BIs to people in need, such as finding time to administer them in busy doctors' offices, obtaining the extra training that helps staff become comfortable providing these interventions, and managing the cost of administering interventions [NIA06]. These challenges may be mitigated by the use of technology.

Patients are often encouraged to use computer programs developed based on BI content in the doctor's waiting room or at home, or to access the interventions through the Internet. Computer-based interventions not only offer privacy, but also the ability to complete the program anywhere, any time of the day [RSB⁺11, WKS⁺10, PSSJC08]. Although computer-based interventions adapted from one-on-one brief interventions are reported to have positive effects on reducing patients' drinking level [RSB⁺11, WKS⁺10, HSD05], they have high drop-out rates because their users loose interest with interacting with the system. One study showed, however, that the delivery of web-based interventions with virtual agents is promising in terms of increasing people's intention to use such an intervention versus an intervention delivered with text only [LAYR13]. That system however is not speech-enabled and the user interacts with mouse and keyboard entries.

Recent developments in the spoken dialogue systems area make deployment of systems that can deliver brief interventions possible. As a natural user interface, spoken dialogue systems may deliver high-quality consistent interventions to screen problem drinkers for alcohol misuse and may encourage users to provide honest disclosure [MF⁺04].

Since brief interventions on alcohol problems are extensively studied and well documented by related institutes [NIA06], I selected the alcohol consumption target behavior. The **objective** of my dissertation project is to create a dialogue system based on machine learning based dialogue management techniques for brief alcohol interventions.

1.1 Statement of the Problem

I posit that the structure and restricted focus of brief interventions (discussed in Section 2.5) make it amenable to computational modeling . The overall problem which my project focuses is development of a proof-of-concept spoken dialogue system that can deliver brief interventions for problem drinkers. The **reinforcement learning** (RL) [SB98] based dialog management approaches will be used which aims at addressing prevalent uncertainty in speech-enabled interfaces. Problems in RL are typically represented based on a mathematical framework called Markov Decision Processes (MDPs) or Partially observable MDPs.

The overall objective of this dissertation project is experimenting and assessing applicability of data-driven, adaptable dialog management approaches [YGTW13, YGK⁺10a, LKW02] in real-world health domain namely for brief interventions for drinking problems. The implications of this research project include, but are not limited to, assessing the feasibility of delivering structured brief health interventions with a data-driven spoken dialogue system, demonstrating reinforcement learning

methods for dialogue management in the health domain, demonstrating the feasibility of modeling relatively long (in terms of number of dialogue turns and number of inputs required to receive) dialogue interaction with both MDPs and POMDPs. In addition, as a result of this research project, there will be a baseline system and a dialogue corpus which can be used in future research in the brief health intervention area. The produced knowledge and experience may also lead to implementation of similarly structured brief health interventions using statistical machine learning methods, for other domains than harmful alcohol consumption (e.g. overeating, lack of exercise, drug use).

I have divided my research problem into four sub-problems. The **first** sub-problem is modelling the brief intervention dialogue with MDPs. MDPs are used for modeling decision making in uncertain situations. Based on the created MDP model, building a baseline system and using its evaluation results as training dialogue corpus is the **second** sub-problem. The **third** sub-problem is optimizing dialogue strategies based on the user population’s interaction experience with the baseline system. The **fourth** sub-problem is remodeling the system with Partially Observable Markov Decision Processes (POMDP) by extending the MDP-based system because it is a more robust framework for uncertainty.

1.2 Specific Research Questions and Objectives

Success of reinforcement learning-based methods in dialogue management illustrates that the approach is also promising outperforms the hand-crafted dialogue management methods [YGK⁺10a, YGTW13]. Therefore, this research project aims at experimenting RL-based techniques in real-world health dialogue application in brief health interventions domain.

The goal of the project is finding solutions for the following **research questions** in brief intervention domain:

- What kind of abstract state representations best serve adaptivity and learning?
- How to optimize initial exploratory state space to avoid data sparsity problem and how to specify appropriate reward function which is needed to learn approximately optimal dialogue strategies?
- How can the large state space be adapted effectively into a tractable subset for reinforcement learning algorithms?
- How to reduce number of exploratory policies –without compromising from optimality of learned policies– for learning approximately optimal dialogue policies with limited number of training dialogue?

To answer the research questions, the following **project objectives** are realized:

Objective 1: In this project, I explore different reinforcement learning techniques and study how they can be applied to the brief intervention dialogue. The overall goal is to **explore possibilities of the state-of-the-art dialogue management techniques to move away from hand-crafted methods employed in most of the current health-related dialogue systems (Discussed in section 2.2), toward data-driven, robust and adaptable dialogue management methods.**

Objective 2:

Statistical dialogue systems that are developed in research projects are usually tested against a baseline in order to make scientific claims. Since there is no baseline dialogue system for brief interventions, one of the objectives is to establish a baseline dialogue system. To reach this goal, I first **design an initial state-based training system which can create exploratory data set.** Despite being exploratory, this system provides desired basic functionality. The baseline system which is compared

with the systems optimized based on reinforcement learning. It helps to assess the effect of optimization objectively.

Objective 3: Using the training dialogue data which is generated by the baseline system from the interactions with real users, I **build an empirical MDP model in which transitions of this MDP models the user population’s reactions and rewards for the various system actions**. This model enables to **compute the optimal dialogue policies based on the MDP model**. I re-test the system based on the learned dialogue policies compare it with the baseline system.

Objective 4: Finally I **change the modelling strategy to POMDPs by extending existing MDPs**. I train the system again and compare it with the baseline system and the MDP-based system. The POMDP-based system is expected to be more robust than the MDP-based system to unseen states.

Some of the **main challenges** of this project are:

- Creating an appropriate representation of each dialogue state by avoiding data sparsity problem.
- Modelling the relatively long (comparing to information seeking dialogue systems) dialogue with MDPs.
- Adapting appropriate reinforcement learning algorithms (e.g. Q-learning) to the learning in large state spaces which brief intervention dialogue create.

1.3 Outline of the Dissertation

In Chapter 2, I explain the background and related work for spoken dialog systems, dialogue management, embodied conversational agents, and brief health interventions. This chapter covers rule-based and machine learning based dialog management methods. It compares and contrasts supervised, and reinforcement learning based

approaches for dialogue management under machine learning based approaches. I cover state of the art dialogue management in health dialogue systems and the recent approaches used in spoken dialogue systems area. Since the system that I develop has an Embodied Conversational Agent interface, I include a survey for them. Finally, I describe the counseling style that my application uses to deliver brief health interventions for problem drinkers.

Chapter 3 describes the approach that is used to develop the system, system components, how I casted brief intervention as Markov decision processes (MDP), and many system design details. In Chapter 3, I described the issues with fully observable MDPs and how can we create a dialogue design tool for health interviews that rely on reinforcement learning based dialogue management approaches. Chapter 5 describes experiments that I conducted with optimized and random dialog policies in terms of subjective and objective aspects of the system.

Chapter 6 covers the tool that is develop to recognize named-entities and domain specific reasoning module for behavioral health. This section describes an ontology design and a system that exploits the created ontology and a semantic network structure to recognize entities and reason about behavioral problems. Chapter 7 is dedicated to sentiment and emotion recognition and it has two sections. First part covers Unified Semantic Network (USN) that is designed to exploit expert knowledge and evaluation of an algorithm that leverages expert knowledge stored in USN to infer about affective ratings and sentiment polarity of words. Second part of Chapter 7 covers sentiment analysis module that can recognize sentiments in short sentences by taking into account sentiments conveyed by named-entities and compositional semantics.

CHAPTER 2

Literature Review

Substantial amount of research has been conducted on developing dialogue systems in restricted microdomains (e.g. flight reservation, information provider systems) in the 90s and early 2000s. Commercial systems have been deployed for reservation/booking, customer center automation, and question answering in the last decade as a result of these early research endeavors [BBC⁺11]. Recently, researchers have mostly concentrated on the development of dialogue systems that are able to take initiative, reason and infer in the domain of the system while conducting dialogue in microdomains using planning, collaborative problem solving [AFB⁺06]. In the last decade, there has been a high interest to develop dialogue systems in real world application areas such as personal coaches, companions and counselors [CDPW08].

While the current trend in commercial systems involves applying finite state-based system directed dialogue strategies, research projects have been moving away from restricted and contrived applications to systems that require flexibility in the management style of the dialogue (e.g. personal coaches, virtual companions) [McT11].

Although there is no accepted standard for dialogue systems design, an end-to-end dialogue system architecture (as shown in Figure 6.2) usually includes three modules: input, output and control [McT02, p. 113]. The input module is commonly composed of Automatic Speech Recognition (ASR) and Spoken Language Understanding (SLU). The control module corresponds to the Dialogue Manager, which executes a dialogue strategy. The output module consists of a Natural Language Generation (NLG) system and a Text-To-Speech (TTS) engine. Depending on the domain of interest

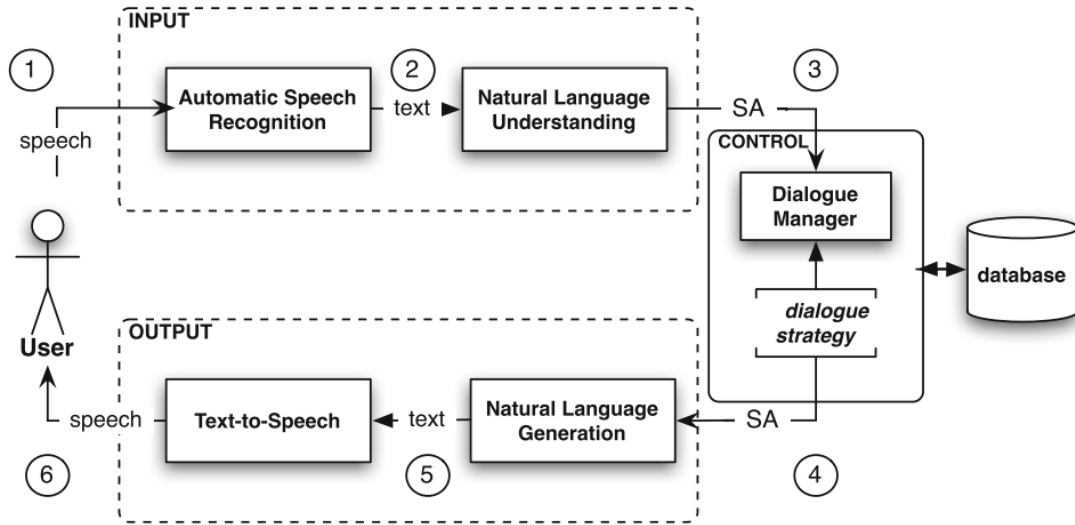


Figure 2.1: Dialogue System Architecture [RL11]

and application area, there can be specialized components such as domain reasoner, planner and specialized databases [FA⁺98].

2.1 Dialogue Management Overview

The primary focus of this research is dialogue management. The Dialogue Manager maintains an internal state and decides what dialogue move (speech act) action to perform next.

There is a wide range of techniques to develop a dialogue manager, and the current techniques used in industry are different from the ones used in research. The focus of this research is dialogue management. The Dialogue Manager maintains an internal state and decides what dialogue move (speech act) action to perform next.

There is a wide range of techniques to develop a dialogue manager, and the current techniques used in industry are different from the ones used in research. Research systems study in aim at conveying natural human behavior such as conversing, reasoning and inferencing, whereas commercial systems aim at being robust to solve a specific

tasks (e.g. customer service automation). Commercial systems are widely based on finite state-based [SCDV⁺98] dialogue management to address their objectives in terms of robustness.

In **finite state-based dialogue management**, dialogue states represent the results of performing a dialogue move from some previous state, and each state leads to a set of allowable next dialogue moves. The information is thus implicit in the state itself and the relationship it plays to other states. While the finite state-based approach is regarded as viable option for simple, scripted dialogue such as reservation systems, it is not possible to create flexible dialogue systems with this approach. Finite state-based approaches have often been compared with plan-based approaches in terms of dialogue modeling .

Plan-based dialog management approaches are viewed as more suitable for flexible dialogue modelling and research in plan-based models is concerned with reasoning domain actions and the goals to which they contribute [FA⁺98, AFB⁺06, BR09, RSL01]. Plan-based approaches are seen as being more opaque than other approaches in terms of development. The plan-based approaches requires a considerable amount of procedural processing power. The plan-based operations (e.g. logical inference) are usually designed based on heuristic rules rather than a well-founded theoretical base. Plan-based dialogue management methods usually involve reasoning and logical inferring operations. Even though, the approach provides a solid computational model, it is not entirely clear for these operations how the mental constructs employed in the model correlate to people's actual mental states.

An **information-state (IS) dialog management** approach allows one to combine the two approaches (finite-state and plan-based), using the advantages of each of them. The information-state may include aspects of dialogue state as well as mentalistic notions such as beliefs, intentions, plans. *“The term information-state of a*

dialogue represents the information necessary to distinguish it from other dialogs, representing the cumulative additions from previous actions in the dialog, and motivating future action” [TL03].

While the information-state approach has been shown to provide a basis for flexible dialogue interaction with rich dialogue state representation, existing IS-based dialogue managers [TL03, BKLO03, tH11] have a number of general limitations that stem from the intuition-based design methodology. From the perspective of system development, the information-state methodology has traditionally been based around a declarative update rule-based design. The functionality of integrating both user dialogue and planning system contributions is encoded as the firing of a complex and sequential update rules. While a rule-based approach is useful in working within an overtly declarative programming environment, the operation of resultant rules can become highly opaque. The IS-based dialogue management, like plan-based approaches, requires the manual specification of update rules that define an action for all possible dialogue situations. It is not practically possible for the designer to anticipate all the possible situations of a dynamic dialogue environment. Thus, dialogue management requires a strategy that can be generalized to unseen events.

2.1.1 Dialogue Management based on Machine Learning

Recent research in dialog management has turned to automated dialog strategy learning using statistical machine learning techniques (e.g. [RKLL11a, YGTW13, YGK⁺10a, TY10]). A major advantage of the **statistical dialog management** approach is that it introduces a principled scientific method for improving dialog strategy design, whereas the previous hand-coded approaches were mainly based on the designer’s intuition.

To date, different Machine Learning methods have been applied [LP⁺07] to automatic dialog management:

- Supervised approaches
- Reinforcement Learning-based approaches

[LP⁺07] has listed key potential advantages of the statistical computational learning approaches for dialog system development as following;

- data-driven development cycle
- provably optimal action policies
- a precise mathematical model for action selection
- possibilities for generalization to unseen states
- reduced development and deployment costs for industry.

Supervised Learning v.s. Reinforcement Learning

The nature of a dialog is *temporal*, and *dynamic*, and machine learning techniques are well positioned to be able address these aspects of dialog.

Dialog is **temporal** in the sense that how good an action is depends on how the dialog progresses further [RL11]. Taking an action affects the state of the dialog and thereby affects the options and opportunities available at later times. Thus, action choice requires foresight and long-term planning with respect to the delayed consequences of actions as specified by to the dynamics of the environment. Therefore, it is not possible to present correct input/output move pairs of ideal dialog strategy behavior. Corpus of dialog usually contains annotations of how good the overall performance of a specific dialog was (e.g. task success, or user scores), but it does not have any indication about how good a single action was. In other words, it is hard

to tell how things should have been exactly done in a particular situation, but it is possible to tell whether the dialog was successful/satisfying overall.

Supervised learning methods do not model dialog as a sequence of actions and they only mimic behavior observed in a fixed dialog corpus. Reinforcement Learning (RL), in contrast, models the problem as sequential decision process with long-term planning [LKW02, RL11]. Therefore, reinforcement learning is more suitable than supervised learning to model the temporal aspect of dialog.

Dialog being **dynamic** describes the fact that dialog takes place in interaction with a stochastic environment, where conditions change frequently (e.g. the level of noise) or a dialog partner reacts differently than predicted in a certain situation. This characteristic requires an approach that is robust to unseen states.

In supervised learning, the learner has to be explicitly instructed what to do in a specific situation by presenting as many examples as possible. RL learns by exploration (of uncharted territory) and exploitation (of current knowledge) [SB98]. The ability to explore allows a system to learn strategies which are more robust to unseen and unpredictable states [RL11, SB98, YGTW13]. The ability to exploit current knowledge allows learning by experience.

2.1.2 Dialogue Management and Reinforcement Learning

Reinforcement Learning (RL) is A sequential decision making, where the RL agent interacts with its environment ([SB98]). The environment is defined as: "anything that cannot be changed arbitrarily by the agent is considered to be outside of it and thus part of its environment" [SB98, p. 53] Reinforcement Learning treats dialogue strategy learning as a sequential optimization problem, leading to strategies that are globally optimal ([SB98]). Uncertainty can be explicitly represented in RL.

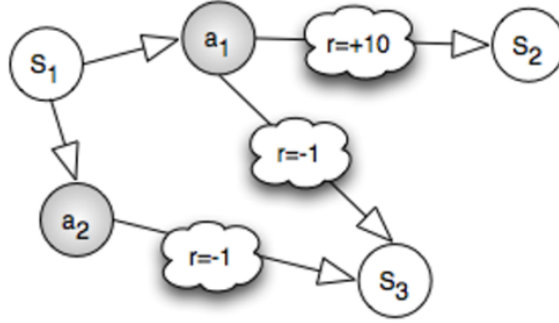


Figure 2.2: RL with Markov Decision Processes (MDPs): The learning agent travels through a network of interconnected states. At each time t , the agent is in state s_t , takes an action a_t , transitions into state s_{t+1} according to the transition probability $p(s_{t+1} | s_t, a_t)$, and receives reward r_{t+1} [RL11].

Within the RL framework for dialogue development, dialogue strategies are represented as mappings from states to actions within Markov Decision Processes (MDPs) [LPE98]. A MDP is formally described by a finite state space S , a finite action set A , a set of transition probabilities T and a reward function R . The dialogue strategy learner can be visualized as an agent traveling through a network of interconnected dialogue states (see Figure 2.2) [RL11]. Starting in some initial state, the learning algorithm transitions from state to state by taking actions and collecting rewards as it goes along. The transitions are non-deterministic, since the dialogue environment is stochastic and dynamic. RL based on MDPs successfully was used by [LKW02, LGHS06, RKLL11b] for implementation of statistical dialogue manager. They tested their system by comparing standard fixed policies and show that the performance of the learned policy is better than several "standard" fixed policies. In a nutshell, the approach followed by [LKW02] in the development of the system is:

- Choosing an appropriate reward measure for dialogues, an appropriate representation for dialogue states, and designing a dialogue policy that maps each state to a set of reasonable actions.

- Building an initial state-based training system that creates an exploratory data set. Despite being exploratory, this system still provides the desired basic functionality.
- Using these training dialogues to build an empirical MDP model on the state space. The transitions of this MDP is modeling the user population’s reactions and rewards for the various system actions.
- Computing the optimal dialogue policy according to this MDP.
- Reimplementing the system using the learned dialogue strategies.

Recently dialogue systems research has become focused on extension of MDP model to handle uncertainty more efficiently which is Partially Observable Markov Decision Processes (POMDP) [YGK⁺10a, YGTW13]. While MDPs assume that the entire state space is fully observable, POMDPs provide a framework for modelling the inherent uncertainty in dialogue systems. POMDPs handle uncertainty by representing the current dialogue state as a belief state which is A distribution over the possible states. The CLASSIC (Computational Learning in Adaptive Systems for Spoken Conversation) project [JKM⁺10] used POMDPs with the information-state approach for dialogue management, which aims to develop a data-driven, machine learning-based architecture for robust and more adaptable spoken dialogue systems (<http://www.classic-project.org>). Some of the results they obtained, are listed below:

- The Hidden Information State system (a POMDP system) improves task success by 25% in high-noise conditions, when tested in simulation [YGK⁺10a]
- 5% reduction in word error rate when using predictions from a simulated user to re-rank n-best lists of speech recognizer’s hypotheses [Lem09]
- Online reinforcement learning improved the commercial application’s completion rate by 10% with real customers [PLL⁺10]

- A statistical planning approach to natural language generation for Information Presentation (content planning and attribute selection) outperforms hand-coded policies and a policy learned from human performance. Tested in simulation in the Tourist Information domain [RLL10]
- Adaptive Natural Language Generation using Reinforcement Learning techniques, evaluation with real users: 12% decrease in time taken, and a 15% increase in task completion rate [JL10]

Reinforcement learning and MDP can be accepted as state-of-the-art techniques for data-driven spoken dialogue system development [YGTW13, FL09]. However, the current systems based on RL and MDPs are usually operating in a tourist information domain where users may ask about hotels, restaurants, activities in a specific location. In this project, I aim at employing these techniques to a brief intervention domain for alcohol. In the next section, I will review current state-of-the-art in health dialogue systems.

2.2 Dialogue Management in Health Dialogue Systems

Although there is a great interest to develop computer-based conversational systems which can promote healthy behaviors or perform interventions, there is very limited experimentation with the state-of-the-art techniques in this area [CDMS07, CD09].

To date there is no system that uses data-driven machine learning-based dialogue management approaches in the health domain. The purpose of this section is to review recent health dialogue systems in terms of dialogue management strategy. Health dialogue systems designed usually based on finite state or rule-based approaches. Therefore, the aforementioned limitations of these approaches (see section 2.1) also valid for the health dialogue systems.

In a recent comprehensive literature review of active assistance technologies in health-related behavior change systems [KPP⁺12], covering articles written between January 2005 and January 2012, 41 health-behavior change systems are identified. Dialog systems and ECAs are identified as emerging technology themes in the behavior change systems field. It is reported that 19 studies out of 41 use dialog systems technology. Among those 19 systems, only 1 system uses speech as an input modality [LL06]. The remaining 19 dialog-based systems use text or menu-based choices as a style of communication. In the same literature review [KPP⁺12], it is reported that 8 of the systems use ECAs [Bic05, BSS11, THSh⁺11, dRNC⁺06].

There are also dialog systems in other health-related domains such as health and fitness companions [THSh⁺11], or virtual support agents for post traumatic stress disorder [MFD⁺12]. Although, there has been growing interest to develop multimodal SDS which can converse, guide, assist or motivate users for different health related topics [BSS11, FQH⁺10, MFD⁺12], to the best of our knowledge, there does not exist any spoken dialog system for the alcohol consumption domain, which is the focus of my research.

Furthermore, dialog management for health-related dialog systems have so far been mostly designed based on *finite state* dialog management mechanisms such as hierarchical transition networks [BSS11, BPS⁺10], plan-based approaches [SBS11b], or information-state based approaches [MFD⁺12]. These systems usually do not have speech recognition integration. Interaction is usually conducted with menu-based choices, but the system utterances are delivered vocally via text-to-speech or prerecorded voice [LAYR13, BSS11].

Other than systems that use menu-based interaction, there are systems that use a different input modality. SimCoach, for example, is a web-based system which uses text as input modality, and an interface with an anthropomorphic ECA which re-

sponds to users' questions with answers prerecorded by a human. Whereas human voices are still superior to synthetic ones, using pre-recorded utterances means that the sentences that the system can speak are fixed, as opposed to systems using text-to-speech engines (like ours) which provides the flexibility of adding new sentences for the system to utter automatically, i.e. without the need to prerecord new sentences. SimCoach, designed to provide support and health-care information about post-traumatic stress disorder, incorporates information-state approach [TL03] with dialog moves with assigned reward values [MFD⁺12]. The Companions project includes three different systems in the domain of health and fitness [THSh⁺11], and all of them use speech as input modality. The dialog management approach in the Companion project is the information-state approach, and one system (the cooking companion) has an ECA interface.

In the computer-based alcohol intervention area, the interaction conducted based on menu-based user interfaces [MGM⁺11, WKS⁺10, HSD05]. To the best of my knowledge, there is no spoken dialogue system which performs alcohol interventions.

2.3 Spoken Dialog Systems

Dialog systems can be classified into two main categories based on their dialog management technique, which can be either based on machine learning (e.g. based on reinforcement learning), or hand-crafted. Systems based on RL are popular in the SDS community and are reported to work better than hand-crafted ones for *speech*-enabled systems [YGTW13, FL09] against noisy speech recognition. *Hand-crafted systems*, on the other hand, can be divided into three subcategories, with dialog management approaches using finite states [SCDV⁺98], plans and inference rules [FA⁺98, BR09] or information states. [TL03].

RL-based dialog systems can learn dialog strategies in a given dialog state from their prior experiences. The idea of having a dialog manager (DM) that can learn interactively from its experience is a cost effective methodology given the alternative approaches: crafting system responses to all possible user’s input using rules and heuristics [PP08]. At best, these rules are based on accumulated knowledge from a trial-and-error experience. At worst, they are based on intuition and limited experience of the designer. Either way, because it is extremely challenging to anticipate every possible user’s input, hand-crafting dialog management strategies is an error-prone process that needs to be iteratively refined and tuned [PP08]. That iterative refinement of course requires substantial amount of time and effort.

The RL-based approach provides the opportunity to automate the design of dialog management strategies by having the system learn these strategies from received reward signals. Potential advantages of statistical dialog management approaches against hand-crafted approaches are listed by [LP⁺07] as 1) a data-driven automatic development cycle, 2) provably optimal dialog action policies, 3) a principled mathematical model for action selection, 4) possibilities for generalization to unseen states, and 5) reduced development and deployment costs.

Approaches for dialog systems based on reinforcement learning (RL) use Markov decision processes (MDP) [SLKW02] or partially observable Markov decision processes (POMDP) [YGK⁺10a, Wil08] frameworks to develop robust dialog managers [FL09, YGTW13]. While both MDP and POMDP require high amount of data for training, POMDPs usually suffer from scalability issues [WY07, YGK⁺10b], and optimization algorithms usually become intractable with large number of states. However, POMDPs are outperforms the MDP-based systems [YGTW13]. It is possible to find tractable solutions by using some approximation to solve practical problems [YGK⁺10b].

In first step of this dissertation project, we used MDP approaches to avoid the mentioned problems associated with POMDPs and to create a baseline for brief intervention domain. Unlike the very classic dialog strategy learning approaches [LPE98, SY02, LPE97] in which the system literally has no knowledge for dialog action selection in the training stage, our system knows taking which actions make sense in each state despite being non-optimal as in [SLKW02]. For example, taking a farewell action in the beginning of dialog instead of greeting does not make sense. Our approach enables our system to learn dialog strategies faster from small amount of dialog corpus than the systems that has absolutely no knowledge in the training. The ideas that are used in [SLKW02] had an influence on our system design decisions, as in NjFun [SLKW02] system we tried to minimize state space and learn dialog policies from real and small amount of data. We extended and adapted some ideas such as state representations and policy design to apply to practical health application.

RL-based dialog systems are mainly used for slot-filling applications. The domain of the dialog is usually in the tourist information domain, such as finding information about restaurants [JTY12, CGP10, YGK⁺10a], appointment scheduling [GWM10], flight reservation [HLG08], or museum guidance [PKKM12]. There has also been work conducted in question-answering [MGLT12] and tutoring domains [CVL10]. The RL-based dialog management paradigm was also recently used for assistive technologies [LGRT13].

In the mentioned systems, each piece of information is accepted as a *slot* which needs to be filled to complete the task. The total number of slots that current systems can usually handle are less than 5, and the flow of a dialog is determined by the slots that need to be filled [SLKW02, YGK⁺10b]. One of the important measures of success of task-based dialogs is their *task completion rate*. When the number of the slots that are needed to be filled increases, the likelihood of successful completion decreases.

I will discuss how our system can handle a substantially larger number of slots than has been attempted so far (we have 18 slots). We will also explain how the flow of the dialog does not depend on unfilled slots, but is rather decided dynamically during the interaction, which is also a novel approach. For example, if our system finds out that the person does not have a drinking problem, it just ends the interaction gracefully. Hence, the length (in terms of slots) and the flow of the dialog are not fixed.

Dialog strategy learning is the main concentration of RL-based dialog systems for the optimization process. Learning optimal dialog strategies can be performed in two ways: based on user simulations [GHL06], or based on interactions with real users in exploration mode [SLKW02, MGLT12]. Since there is no data in our domain, we had to follow the latter way. It is worth mentioning that simulated users can have some shortcomings to realistically reproduce the behaviors of the diverse real users' behaviors [SGY05].

2.4 Embodied Conversational Agents and Spoken Dialog Systems

Virtual human-like characters that specifically focus on dialog-based interactions are called **Embodied Conversational Agents (ECAs)**, also known as Intelligent Virtual Agents (IVA). ECAs are digital systems created with an anthropomorphic embodiment (be it graphical or robotic), and capable of having a conversation (albeit still limited) with a human counterpart, using some artificial intelligence broadly referred to as an “agent”. With their anthropomorphic features and capabilities, they interact using humans' innate communication modalities such as facial expressions, body language, speech, and natural language understanding, and can also contribute

to bridging the digital divide for low reading and low health literacy populations, as well as for technophobic individuals [NK11, BPJ09].

One of the most influential work for the study of virtual animated characters established that, when provided with social cues by a computer system, humans react socially similarly to how they would with a human [RN96]. Because latest ECAs can use their sophisticated multimodal communication abilities to *establish rapport* [Pel09, KGWW08a, KGWW08b, WG10, WG09, GOL06, GWGF07, PKG, MRP08, PI05, HMG11], *communicate empathically* [PI05, ML07, AVSP07, BBA07, BWHG09, BW11, AVSP07], and *engage in social talk* [Klu11, BGP05, Bic05, BG06, CB03, SBS11a], they have become capable of being as engaging as humans - and have even been found more engaging than humans at times [GWO07]. In this dissertation, the developed spoken dialog systems have an ECA interface.

Intelligent virtual agents (IVA) - also known as embodied conversational agents (ECA) or virtual humans (VH) - and spoken dialog systems (SDS) are two emerging fields of research which, *together*, could bring a revolution to human-computer interaction as we know it. Even though the term ECA includes the notion of spoken dialog, SDS and ECA communities still do not have a strong connection. While progress in the spoken dialog system area is complementary for the development of conversational embodied agents, latest findings in SDS research have not been commonly used by ECA researchers (and vice versa).

Indeed, although spoken dialog systems (SDS, henceforth) research has shown in the past few years that using Reinforcement Learning (RL) with MDPs for dialog management outperforms older hand-crafted rule-based approaches [FL09, YGTW13], intelligent virtual agent researchers have not yet integrated these results in their dialog systems. ECA-based systems usually involve spoken dialog (versus menu options



Figure 2.3: Multimodal Embodied Conversational Agent Interface

to choose from), but their dialog management usually still relies on hand-crafted methods [MDS⁺12, BG10].

In this project, I bring together latest progress from the SDS community to the IVA community with the use of RL-based dialogue management integrated with a 3D animated character (shown in Figure 2.3). The 3D animated virtual character is an interface for a task-based spoken dialogue to deliver brief alcohol interventions to people at-risk of health issues due to excessive alcohol consumption.

From a computer science perspective, my work aims at building a fully implemented system to be used as screening tools to help individuals at risk of health issues, and at evaluating the system in terms of both, users' (subjective) acceptance and dialog system's (objective) performance.

From a health-care perspective, I aim at increasing access to effective evidence-based health interventions with a novel mode of delivery for computer-based health interventions - namely delivering health interventions with a virtual counselor. The screening dialog system brings insight and awareness regarding *alcohol problems* by using the well established *brief intervention* (BI) counseling approach. BIs are short, well structured, one-on-one counseling sessions, focused on specific aspects of problematic lifestyle behavior. BIs are not only ideally suited for people who drink in ways that are harmful or abusive (which is the current domain of our work), but BIs have also been used successfully for a variety of target problem behaviors (e.g. overeating, lack of exercise). Therefore the results of our research will also have an impact on dialog systems for diverse behavior change interventions for healthy lifestyles.

2.5 Brief Interventions for Alcohol Problems

Unlike traditional alcoholism treatment, which focuses on helping people who are dependent on alcohol, brief interventions or short, one-on-one counseling sessions are ideally suited for people who drink in ways that are harmful or abusive [MFSV02]. Brief interventions can be delivered in few minutes and require minimal followup whereas traditional alcoholism treatment takes many weeks or months.

The purpose of brief interventions is different from formal alcoholism treatment. Brief interventions generally aim to moderate a persons alcohol consumption to reasonable levels and to eliminate harmful drinking behaviors (such as binge drinking), rather than to insist on complete avoidance from drinking – although abstinence may be encouraged, if appropriate [MF⁺04]. Reducing levels of drinking or changing patterns of harmful alcohol use helps to reduce the negative outcomes of drinking, such as alcohol-related medical problems, injuries, domestic violence, motor vehicle crashes, arrests, or damage to a developing fetus.

Brief interventions typically consist of one to four short counseling sessions with a trained interventionist (e.g., physician, psychologist, social worker). [MFSV02] looked at 34 different studies and found that people who received brief interventions when they were being treated for other conditions consistently showed greater reductions in alcohol use than comparable groups who did not receive an intervention. People seeking treatment specifically for alcohol abuse appeared to reduce their alcohol use about the same amount, whether they received brief interventions or extended treatments (five or more sessions). These findings show that brief interventions can be an effective way to reduce drinking, especially among people who do not have severe drinking problems requiring more intensive treatment.

Brief interventions may include approaches such as **motivational interviewing** [MR02]. Motivational interviewing is “a collaborative, person-centered form of guiding to elicit and strengthen motivation for change” [MR⁺09a]. It is designed to elicit intrinsic motivation of people who are resistant to moderating their alcohol consumption or who do not believe they are drinking in a harmful or hazardous way. Motivational interviewing [MR02] encourages patients to decide to change for themselves by helping to resolve their ambivalence about change. Clinicians also can assist patients by helping them establish specific goals and build skills for modifying their drinking behavior.

Delivering Brief Intervention

Brief interventions give patients a simple way to receive care in a comfortable and familiar setting. Because they are brief, they can be easily incorporated into a variety of medical practices. Moreover, these approaches offer a lower cost alternative to more formal alcoholism treatment [MF⁺04].

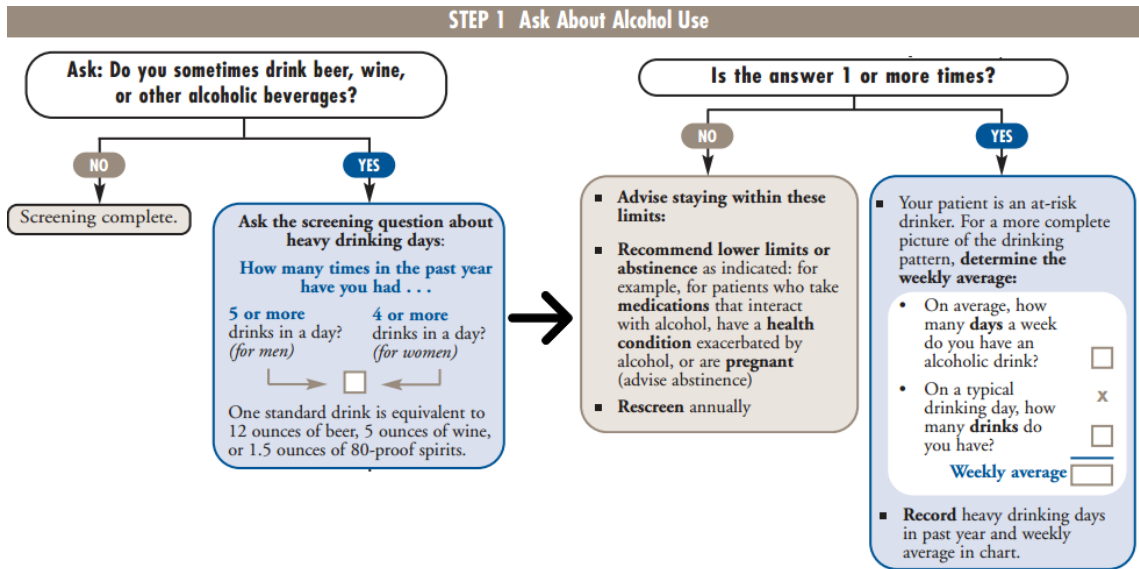


Figure 2.4: First Step of Alcohol Screening and Brief Intervention [NIA06]

Typically a nonspecialist staff who the patient may already trust or feel comfortable being treated by such as a physician, a nurse, or physicians assistant in a primary care setting, or nurse or physicians assistant on a medical unit - delivers the brief intervention [MF⁺04].

Supplemental handouts may be provided to patients during the intervention, including pamphlets, manuals, or workbooks to reinforce the strategies offered during the session. Clinicians also can follow up at a later date, either in person or through the mail, to provide additional assessment and further motivate the patient to achieve the goals set during the initial meeting. If the brief intervention does not motivate the patient to reduce alcohol consumption, clinicians can recommend more intensive treatment.

Many of the challenges involved in administering brief interventions - such as finding the time to administer them in busy doctors offices, obtaining the extra training that helps staff become comfortable providing interventions, and managing the cost of using interventions - may be overcome through the use of technology. Patients are

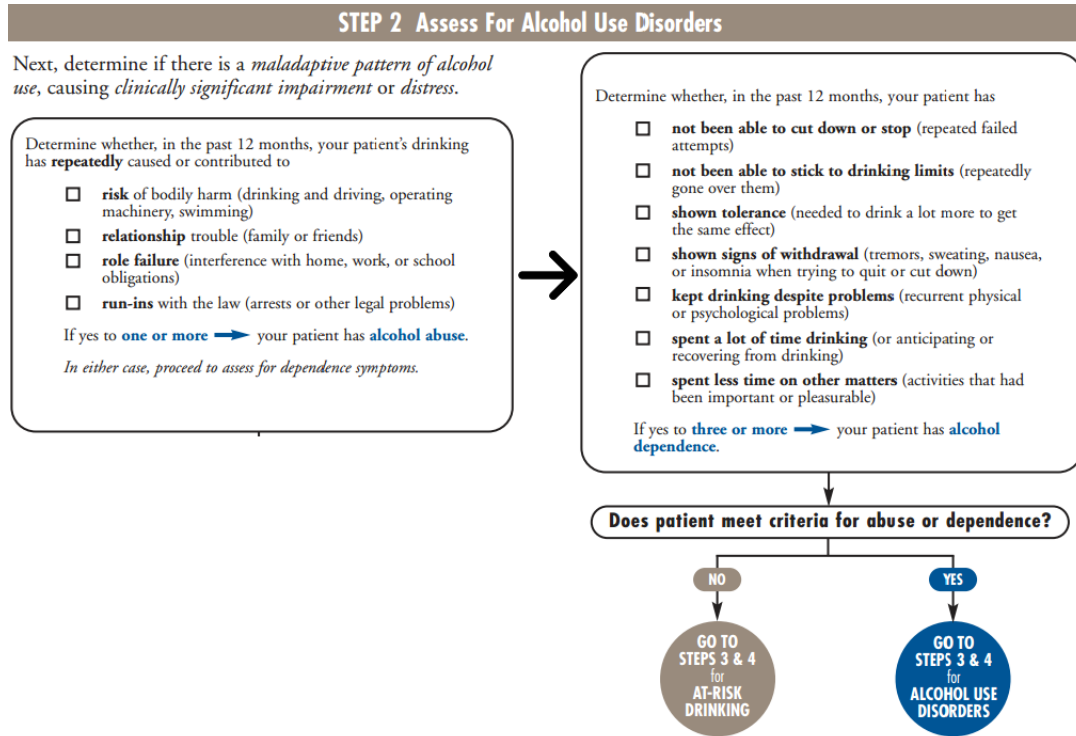


Figure 2.5: Second Step of Alcohol Screening and Brief Intervention [NIA06]

sometimes encouraged to use computer programs in the doctors waiting room or at home, or to access the intervention through the Internet, which offers privacy and the ability to complete the program at any time of day [HSD05, PSSJC08, MGM⁺11].

The current research is focused on to delivering interventions by incorporating spoken dialogue technology with anthropomorphic user interfaces such as interactive virtual characters [BSS11, SBS11b, CDPW08]. Pilot results indicate that although users reported they would be most comfortable consulting with a doctor in person [GBSM⁺03], they responded positively to the computerized interventions [Bic05, SBS11b, BPJ09], which were accessible even to those with little computer experience.

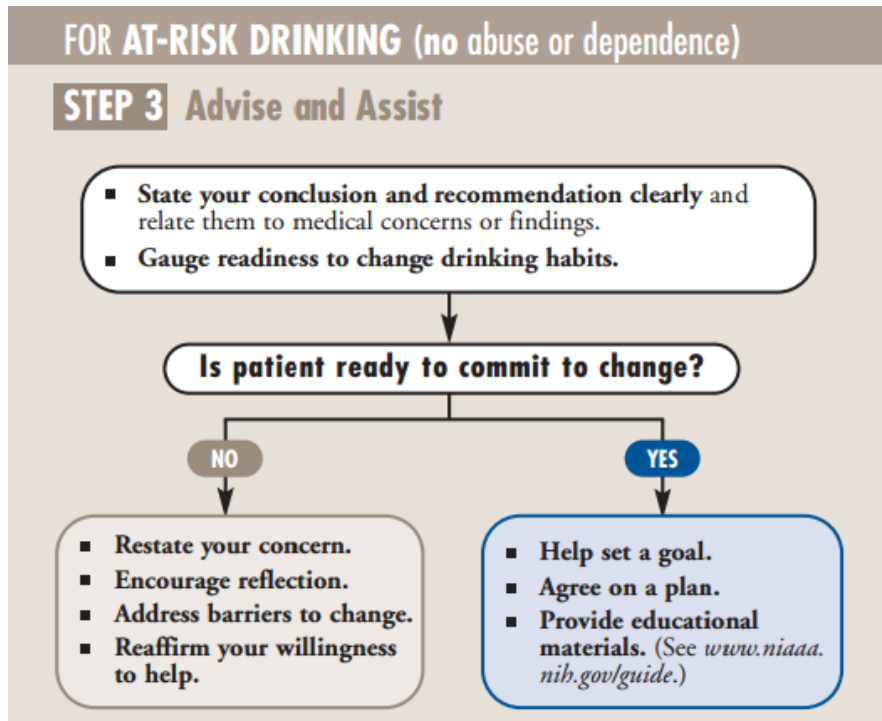


Figure 2.6: Third Step of Alcohol Screening and Brief Intervention for At-Risk Drinking [NIA06]

Brief Intervention Interaction For Alcohol

According to the clinician's guide for conducting brief interventions [NIA07a] from National Institute on Alcohol Abuse and Alcoholism (NIAAA), a brief intervention can be delivered in three steps;

- Step 1: Asking About Alcohol Use (See Figure 2.4)
- Step 2: Assessing for Alcohol Use Disorders (See Figure 2.5)
 - Assessment of Abuse
 - Assessment of Dependence
- Step 3: Advising and Assisting according to degree of alcohol problem
 - At-risk drinkers (See figure 2.6)

– Drinkers with alcohol use disorder (See figure 2.7)

The guide provides two methods for the first step (See Figure 2.4) ; a single question (about heavy drinking days) or administering a written self-report instrument. The single interview question can be used at any time, either in conjunction with the screening instrument or alone. For this project, the system start intervention with a single question about alcohol use (e.g. "Do you sometimes drink beer, wine, or other alcoholic beverages?"). If the client's answer is no, there is no need to continue to screening. If the client's answer is yes, the system will ask about amount of the alcohol client consumes to find out if the client is at-risk drinker (e.g. "How many times in the past year have you had 5 or more drinks in a day").

If a client is not an at-risk drinker, the system may advise maintaining or lowering drinking limits according to the situation and offer re-screening annually. If a client is an at-risk drinker, to get the complete picture of drinking, the system will ask several questions to query the drinking pattern of a client (e.g. On average, how many days a week do you have an alcoholic beverage?).

In step 2 (See Figure 2.5), the system will try to determine whether or not there is a maladaptive pattern of alcohol use that causing clinically significant impairment or distress. In this step, the system will try to query if a client has alcohol abuse (e.g. risk of bodily harm, relationship trouble) and alcohol dependence (e.g. kept drinking despite problems, not been able to stick to drinking limits) problem. If a patient does not meet the criteria for alcohol abuse or dependence, the patient is still at-risk for developing alcohol related problems. If a patient has an alcohol use disorder (dependence or abuse), the next step (step 3) will be different than at-risk drinkers.

In step 3, if the client is at-risk (See figure 2.6), the system will state its conclusion according to the guideline and recommend to cut down his/her drinking. Then it will

FOR ALCOHOL USE DISORDERS (abuse or dependence)

STEP 3 Advise and Assist

- **State your conclusion and recommendation clearly** and relate them to medical concerns or findings.
- **Negotiate a drinking goal.**
- **Consider evaluation by an addiction specialist.**
- **Consider recommending a mutual help group.**
- For patients who have dependence, **consider**
 - the need for **medically managed withdrawal** (detoxification) and treat accordingly.
 - prescribing a **medication** for alcohol dependence for patients who endorse abstinence as a goal.
- **Arrange followup appointments**, including medication management support if needed.

Figure 2.7: Third Step of Alcohol Screening and Brief Intervention for Alcohol Use Disorders [NIA06]

try to assess readiness to change, if the client is not ready to change, the system will restate its concern for client's health, encourage reflection by asking positive versus negatives of drinking and reaffirm its willingness to help when the client is ready. If a client is ready to change, the system will set a goal, agree on a change plan and provide educational materials (e.g. pamphlets). In step 3 for the clients who has alcohol abuse or dependence problems (See figure 2.7), the system will state its conclusion, negotiate drinking goal and refer to an addiction specialist. A sample brief intervention dialogue showed in Table A.4. The system performs screening in the first step with 3 questions. In the second step, it asks a question for assessing alcohol abuse. Since it elicits indicator of alcohol abuse, it does not ask about alcohol dependence (one indicator is enough for abuse) and passes to the next step. Based

on the result from the step 2, the system delivers step 3 which is for people who has alcohol use disorders. It recommends to see to get professional help.

Table 2.1: Sample Dialog: S shows systems utterance and C shows user utterance in Turn Column. Second column shows what user or system utter in each turn. Third column shows in which step the system is operating in terms of Brief Intervention.

Turn	Utterance	Step in BI	
S1:	Hi, My name is Lola, I am a virtual character created to have conversations with you and to teach you about healthy lifestyles. Today, I'm here to talk about drinking to figure out if you have any issues. May I ask you few questions about your drinking?		
C1:	Yes, sure you can.		
S2:	Do you sometimes drink beer, wine, or other alcoholic beverages?		
C2:	Well yes, I drink wine or beer sometimes!		
S3:	How many times in the past year have you had 5 or more drinks in a day?		
C3:	I think around once a week.		
S4:	Actually, consuming more than 4 drinks on any single day is classified as heavy drinking. So it seems like you are sometimes going above heavy drinking limits? Is that right?		
C4:	Yes, that's correct.		
S5:	How frequently do you have an alcoholic beverage each week?		Step 1: Screening
C5:	I think at least 3 days a week.		
Continued on next page			

Table 2.1 – continued from previous page

Turn	Utterance	Step in BI
S6:	On a typical drinking day, how many drinks you have?	
C6:	I just drink... maybe a bottle of whisky.	
S7:	Sorry, I could not understand your answer, can you please repeat?	
C7:	I said one bottle of scotch.	
S8:	Ok, thanks for the information you have provided me about your drinking. Next I will try to determine some important effects of your drinking.	
S9:	Have you ever taken risks under the influence, such as drinking and driving, operating a motor vehicle, or swimming?	
C9:	No I have not.	
S10:	I see, so you are careful to avoid high risk situations after or while having drinks, right?	
C10:	Yes, I try to be careful.	Step 2: Abuse
S11:	Have you continued to drink even though you knew it was causing you trouble with your family or friends?	Assessment
C11:	I think yes, I continue to drink while my family complains about my drinking.	
S12:	Thanks for the information you've provided. In the next step I will ask you questions about potential dependence-related aspects of your drinking. Is that okay?	
C12:	Yes I want to know, please go ahead.	
Continued on next page		

Table 2.1 – continued from previous page

Turn	Utterance	Step in BI
S13:	Have you had times when you ended up drinking more than you meant to? Or kept on drinking for longer than you intended?	
C13:	Not usually, but... yes I guess I did.	
S14:	Have you found that you have to drink much more than you once did in order to get the effect that you like?	
C14:	I think my tolerance is increased.	
S15:	Sorry, I could not understand your answer. Do you need to drink a lot more sometimes to get the same effect?	
C15:	I meant yes, I need to drink more and more.	
S16:	Ok sorry but just be sure: you can not get the same effect with the same amount of alcohol. Is that right?	
C16:	Yes, that is correct.	
S17:	Have you more than once wanted to stop or cut down on your drinking? Or tried more than once to stop or cut down but found you couldn't?	
C17:	Well, I think I tried but I was not that serious about stopping or cutting down.	Step 2:
S18:	I see, so you really never attempted to stop or cut down. Is that right?	Dependence
C18:	No, it is wrong.	
S19:	Sorry for the misunderstanding, can you please repeat your answer? Have you more than once wanted to stop or cut down on your drinking but found you couldn't?	
Continued on next page		

Table 2.1 – continued from previous page

Turn	Utterance	Step in BI
C19:	Yes I did but not seriously.	
S20:	Ok, now, when the effects of alcohol are wearing off, have you had trouble sleeping, or found yourself shaking, nervous, nauseous, restless?	
C20:	I found myself sweating and my heart beating fast.	
S21:	Sorry I could not understand your answer. Can you please repeat it?	
C22:	I said I found myself sweating and my hard beating fast.	
S22:	I am sorry for asking again, can you repeat your answer please.	
C22:	I said I am sweating and my heart beating fast when I drink too much.	
S23:	Have you had a period when you spent a lot of time drinking? Or being sick or getting over the bad aftereffects of drinking?	
C23:	Yes several times.	
S24:	Thank you very much for all your answers. Now I'd like to just take one or two more minutes of your time. Is that okay?	
C24:	Yes, please go ahead.	
S25:	On a scale from 1 to 10, how ready are you to make a change?	
C25:	I think 7	Step 3:
Continued on next page		

Table 2.1 – continued from previous page

Turn	Utterance	Step in BI
S26:	Great! Just for your information, men who drink more than 4 standard drinks in a day or more than 14 per week, and women who drink more than 3 drinks in a day or more than 7 drinks per week, are at increased risk of alcohol related problems. Based on your answers, I recommend you to stick with this limit or even abstain if you can. I also would like you to track your drinking, and I will send you a calendar for that. I'm also going to give you information about places near you where you can get help if you choose to. Finally, I would like to talk with you again 3 month later. Is it okay?	Disorder
S26:	Yes, it's okay.	
S27:	Thanks for talking with me. I hope you've learned something useful about your drinking pattern. Good Bye and talk to you in 3 months.	

CHAPTER 3

Spoken Dialogue System Based on Markov Decision Processes

In the first phase of the spoken dialog systems project, I use Markov Decision Processes (MDP) for dialog management. In this section, I discuss details of the system. The most of the system components that are described in this section are used throughout this research project except the dialog management methodology. The dialog management methodology in this section is based on MDPs, in the next section it is based on partially observable MDPs.

3.1 Approach

An overview of our system architecture for MDP-based system is shown in Figure 8.1, and explained in details in the Section 3.5. In short, we use reinforcement learning for the dialog management, combined with a 3D animated character who converses with the user with Text-To-Speech (TTS) utterances (versus fixed prerecorded sentences). The domain is a task-based spoken dialog to deliver brief alcohol interventions to people, and identify whether they are at-risk of health issues due to excessive alcohol consumption.

According to the clinician’s guide for conducting brief interventions from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) [NIA07a], a brief intervention for alcohol-related health problems can be delivered in three sequential steps:

- Step 1: *Asking* About Alcohol Use
- Step 2: *Assessing* for Alcohol Use Disorders
 - Assessment of Abuse
 - Assessment of Dependence
- Step 3: *Advising* and Assisting according to degree of alcohol problem

- At-risk drinkers
- Drinkers with alcohol use disorder

To develop the dialogue content, we follow the brief intervention guide for alcohol prepared by NIAAA [NIA06]. The details of brief intervention for alcohol is discussed in Section 2.5, but I briefly review it in this section. The goal of the dialogue system is to deliver alcohol screening and brief interventions based on this guide. Each step contains a set of questions.

In *Step 1*, there are 5 questions. The system asks these 5 questions, and if the user expresses that s/he is not consuming alcohol from time to time, the interaction is gracefully terminated by the system. Otherwise, the dialog manager continues to the second step.

In Step1, there are 5 questions. The system asks these 5 questions, and if the user expresses that s/he is not consuming alcohol from time to time, the interaction is gracefully terminated by the system. Otherwise, the dialog manager continues to the second step.

In *Step 2*, in the Assessment of *abuse* stage, there are 4 questions to assess alcohol abuse indicators. It is enough to find one indicator of alcohol abuse (e.g. risk of bodily harm, relationship trouble) to move to the Assessment of *dependence* stage (e.g. keep drinking despite problems, not able to stick to drinking limits). If the system can not find any indicator of abuse with the 4 questions, it passes to the dependence stage. In the Assessment of *dependence* stage (still in *Step 2*), there are 7 questions.

In Step 2, in the Assessment of Abuse stage, there are 4 questions assessing alcohol abuse indicators. It is enough to find one indicator of alcohol abuse to move to the assessment of dependence stage. If the system can not find any indicator of abuse with the 4 questions, it passes to the dependence stage. In the dependence stage, there are 7 questions.

It is enough to detect 3 dependence indicators to transit to *Step 3*, Advice for drinkers with alcohol use *disorder*. If the system does not detect 3 dependence indicators, it transits to Advice for *at-risk* drinkers. Therefore, the dialog branches to two separate steps in *Step 3*: **1)** one for *at-risk* drinkers, and **2)** one for drinkers with alcohol use *disorder*. In both branches, the system provides information related to the assessment of the system. If the system assessed that the user has an alcohol use disorder, it refers the user to treatment, asks the user if she or he is ready to change, and suggests a goal toward a change of drinking patterns, based on the user’s readiness. If the user is an at-risk drinker, it gauges his or her readiness to change, and provides feedback and information about the person’s drinking. Therefore in both stages, the system provides factual information about the person’s drinking and suggested drinking limits, and asks what is the user’s intention to change with a single question. In total there can be a maximum of 18 different questions in a single session.

A sample dialog between the system and the user is shown in Table A.4. The dialog we presented in the table covers em Step 1 Screening about alcohol *use*, and *Step 2* Assessment of *abuse* completely. In *Step 2* Assessment of *dependence*, there are 7 questions, and then the system branches to Step 3. Because of space limitations, we cannot present a longer dialog. The system uses questions recommended by NIAAA. It uses simple reflections for confirmations instead of explicit confirmations, and examples are shown in Table A.4 with the S4 and S10 system dialog turns.

3.2 Reinforcement Learning Background

I built the system using reinforcement learning. *Reinforcement Learning* (RL, henceforth) is a sequential decision making algorithm, where the RL agent interacts with its environment [SB98]. The environment is defined as: “anything that cannot be

changed arbitrarily by the agent is considered to be outside of it and thus part of its environment” [SB98, p. 53]. Reinforcement learning treats the learning of dialog strategies as a sequential optimization problem, leading to strategies which are globally optimal [SB98].

Within the RL framework for dialogue development, dialogue strategies are represented as mappings from states to actions within Markov Decision Processes (MDPs) [LPE98]. In other words, a dialog strategy specifies, for each system state, what is the next action to be taken by the system [LPE98, SY02]. The MDP framework can be characterized by a 4-tuple (S,A,T,R), where:

- S is a finite set of states
- A is a finite set of actions
- T is a state-transition function such that $T(s, a, s') = \mathbf{P}(s' | s, a)$ which describes how the probability of performing action **a** in state **s** will lead to state **s'**
- $\mathbf{R}(s, a, s')$ is a local reward function such that $R(s, a) = \sum_a \mathbf{P}(a|s)\mathbf{R}(s, a, s')$, and the objective of the SDS is to maximize the gained reward.

3.3 Markov Decision Processes

To develop the system, I needed to represent the internal states of the system using MDP formalization as described earlier. States, actions and state transitions needed to be created. First, I elicited the state attributes. I considered the most important dialog attributes to represent the dialog state concisely. To avoid the data sparsity problem during training, I divided the whole system into 5 sections according the BI guide steps. Hence, there are 5 MDPs in the system. For the each step, I elicited related state features.

For each state, there are 5 common **attributes**: Question, Confidence, Value, Grammar, and Aux (see Table 3.1 for the descriptions of each attribute). There is usually 1 extra attribute to represent the step-specific requirement at each step, e.g. Greet because in the first step the system needs to greet the user first and get consent of the user to start the dialog. The full list of state features for *Step 1* are listed in Table 3.1.

I then designed possible states: for each question there are 34 possible states. Each state represents the state of the conversation. For example, dialogue state *112101* indicates that the system has greeted the user (G=1), the first question has been queried (Q=1), the ASR confidence level is high (C=2), the value is obtained (V=1), the type of grammar is restrictive (Gr=0) and that the polarity of the alcohol usage indicator (i.e. positive or negative indicator such as consuming alcoholic beverages above limits is positive, below limits is negative) is positive (Ax=1). The Aux feature is set to 0 when it is not used. It is used for multiple purposes such as keeping track of the number of re-asked questions or of the polarity of alcohol problem indicator.

Table 3.1: State attributes and values for Step 1: Asking About Alcohol Use

Attribute	Values	Description
Greet (G)	0,1	Whether the system has greeted the user
Question (Q)	1,2,3,4	Which question is being queried
Confidence (C)	0,1,2,3, 4,5,6	0,1,2 for low, medium, and high confidence of speech recognizer. 3,4 for confirmed or not confirmed. 5 to indicate system is waiting for confirmation. 6 is for to indicate system transit to next question without confirmation
Value (V)	0,1	Is the value obtained for current question
Grammar (Gr)	0,1	What type of ASR grammar used, restrictive or dictation (non-restrictive) grammar.
Aux (Ax)	0,1,2	Multiple purpose attribute. Use to indicate number of ReAsks and semantic valence of the received answer. If it is 0, it indicates, it is not used in that state.

Table 3.2: Dialog Actions For Step 1 Question1

Dialog Action	System Utterances
S1-AskQ1Sys:	Do you sometimes drink beer, wine, or other alcoholic beverages?
S1-AskQ1User:	Can you briefly talk about your alcohol consumption?
S1-ReAskQ1Sys	Sorry, I could not understand your answer. Do you sometimes drink beer, wine, or other alcoholic beverages?
S1-RaAskQ1User:	Sorry, I could not understand your answer. Can you briefly talk about your alcohol consumption?
S1-ConfQ1Pos	So you like to have alcoholic beverages time to time, is that right
S1-ConfQ1Neg	so you are recently not having any alcoholic beverages, is that right
S1-NoConf	—
S1-NotConfirmedQ1Sys	I am sorry for the misunderstanding do you sometimes drink beer, wine, or other alcoholic beverages?
S1-NotConfirmedQ1User	I am sorry for the misunderstanding, can you briefly talk about your alcohol consumption?

We refined our state representations by excluding the states that make no sense and manually checking each state. For example, state 040111 is a non-sensible state in *Step 1*. The system can not be in that state because the system needs to greet the user *first*, before it can ask questions. It is not possible to ask question 4 in the first step without greeting or without asking prior questions. Excluding non-sensible states yielded a very large state-space reduction.

For *Step 1 Use*, *Step 2 Abuse*, *Step 2 Dependence*, *Step 3 At-risk*, and *Step 3 Disorders*, the number of states are 170, 136, 238, 68 and 68 respectively, i.e. 34 states multiplied by the number of questions in a step). Total number of states is 680.

I then created our **dialog actions** for each question. The system uses 2 types of initiative dialog actions: system initiative, where the system asks close-ended questions (e.g. Do you sometimes drink beer, wine, or other alcoholic beverages?), and user initiative, where the system asks open-ended questions (e.g. Can you

briefly talk about your alcohol consumption?). We refer to the system/user initiative terms in the same manner as they are widely referred to in the SDS community [SLKW02, SKL⁺00]. In the system initiative questions (where the expected answer is relatively restricted given the close-ended nature of the questions), the system uses a restrictive grammar for speech recognition (SR). In the user initiative, the system uses a non-restrictive SR grammar to handle user's answer to open-ended questions.

There are 9 possible actions for each question, which are grouped under 4 categories: **1)** *Ask* actions are used when the system needs to ask a question to the user for the first time, which can be performed with the two types of initiatives; **2)** *ReAsk* actions are used if the system can not understand the user's speech, which can also be done with two types of initiatives; **3)** *Confirmation* actions are used to ask for confirmation as to whether the system understood what the user said.

The confirmation actions are system initiative by default, and are of 3 types: positive, negative and no confirmation. The positive type is used if the system receives an answer which reveals alcohol usage/abuse/dependence information. The negative type is used if the system receives an answer which indicates no drinking/abuse/dependence problem. The third type is no-confirmation, and it is used if the system decides to pass to the next question without confirmation (possible action at any question). *NotConfirmed* action is used if the user gives a negative answer to a confirmation action.

The available actions for the first question in *Step 1 Screening About Alcohol Use* are shown in Table 3.2. The first column is the name of the dialog action, and the second column is the system utterance. The actual name of the dialog action starts with the step information (e.g. S1), then the type of dialog action (e.g. Ask), then the question being queried (Q1) and the initiative type (Sys). So S1-AskQ1Sys stands

for *Step1* (S1), the question type is *Ask*, the question being queried is *question one* (Q1) and the initiative type is *system initiative* (Sys).

The number of available actions for each question is 9 (as for the first question shown in Table 3.2). Although the length of the dialog is not fixed, our system asks a maximum of 18 questions. There are 162 available actions (for asking questions, re-asking questions, and confirmations) for the system to select from in the longest dialog session (18 questions multiplied by the number of available actions). There are dialog actions which are used while transiting from one step to another step (e.g. from *Step 1* to *Step 2 Abuse*) and dialog actions for ending the conversation. There are 2 actions for giving feedback to at-risk drinkers and to drinkers with alcohol use disorder at the end of the each session. The total number of the dialog actions is 169.

After creating the dialog actions, I created the **dialog policies**. A dialog policy is a mapping of a state to sensible dialog actions. I mapped each state to 2 possible dialog actions based on the initiative or confirmation type. Table 3.3 shows exploratory dialog policies for Question 1 in Step 1.

As we mentioned earlier, for each question there are 34 states. State updates are performed based on user's dialog actions or on systems dialog actions in each dialog turn. In Table 3.3, only 30 state-actions mappings that are updated by the system dialog actions or user dialog actions are shown. The remaining 4 states are only updated based on user's dialog actions, which is why we did not include them in Table 3.3. The reason for this is that, if the system waits for the confirmation from the user (i.e. where $C=5$ as shown in see Table 3.1), the system dialog actions can not be used to update a state. In other words, the remaining 4 states need to be updated by user's dialog actions. In Table 3.3, we only show the states that are updated by the system. However, the states in Table 3.3 are the result of the user's dialog actions since *Value Grammar*, *Confidence* and sometimes *Aux* are updated

Table 3.3: Explanatory Policies For Step 1 Question 1

States						Available Actions per State
G	Q	C	V	Gr	Ax	
1	1	0	0	0	0	S1-ReAskQ1Sys, S1-ReAskQ1User
1	1	0	0	1	0	S1-ReAskQ1Sys, S1-ReAskQ1User
1	1	0	0	0	1	S1-ReAskQ1Sys, S1-ReAskQ1User
1	1	0	0	0	2	S1-AskQ2Sys, S1-AskQ2User
1	1	0	0	1	1	S1-ReAskQ1Sys, S1-ReAskQ1User
1	1	0	0	1	2	S1-AskQ2Sys, S1-AskQ2User
1	1	0	1	0	1	S1-ConfQ1Pos, S1-NoConf
1	1	1	1	0	1	S1-ConfQ1Pos, S1-NoConf
1	1	2	1	0	1	S1-ConfQ1Pos, S1-NoConf
1	1	0	1	1	1	S1-ConfQ1Pos, S1-NoConf
1	1	1	1	1	1	S1-ConfQ1Pos, S1-NoConf
1	1	2	1	1	1	S1-ConfQ1Pos, S1-NoConf
1	1	0	1	0	2	S1-ConfQ1Neg, S1-NoConf
1	1	1	1	0	2	S1-ConfQ1Neg, S1-NoConf
1	1	2	1	0	2	S1-ConfQ1Neg, S1-NoConf
1	1	0	1	1	2	S1-ConfQ1Neg, S1-NoConf
1	1	1	1	1	2	S1-ConfQ1Neg, S1-NoConf
1	1	2	1	1	2	S1-ConfQ1Neg, S1-NoConf
1	1	6	1	1	1	S1-AskQ2Sys, S1-AskQ2User
1	1	6	1	1	2	S1-Q1End, S1-Q1End
1	1	6	1	0	1	S1-AskQ2Sys, S1-AskQ2User
1	1	6	1	0	2	S1-Q1End, S1-Q1End
1	1	3	1	0	1	S1-AskQ2Sys, S1-AskQ2User
1	1	3	1	1	1	S1-AskQ2Sys, S1-AskQ2User
1	1	3	1	0	2	S1-Q1End, S1-Q1End
1	1	3	1	1	2	S1-Q1End, S1-Q1End
1	1	4	1	0	1	S1-NotConfirmedQ1Sys,S1-NotConfirmedQ1User
1	1	4	1	1	1	S1-NotConfirmedQ1Sys,S1-NotConfirmedQ1User
1	1	4	1	0	2	S1-NotConfirmedQ1Sys,S1-NotConfirmedQ1User
1	1	4	1	1	2	S1-NotConfirmedQ1Sys,S1-NotConfirmedQ1User

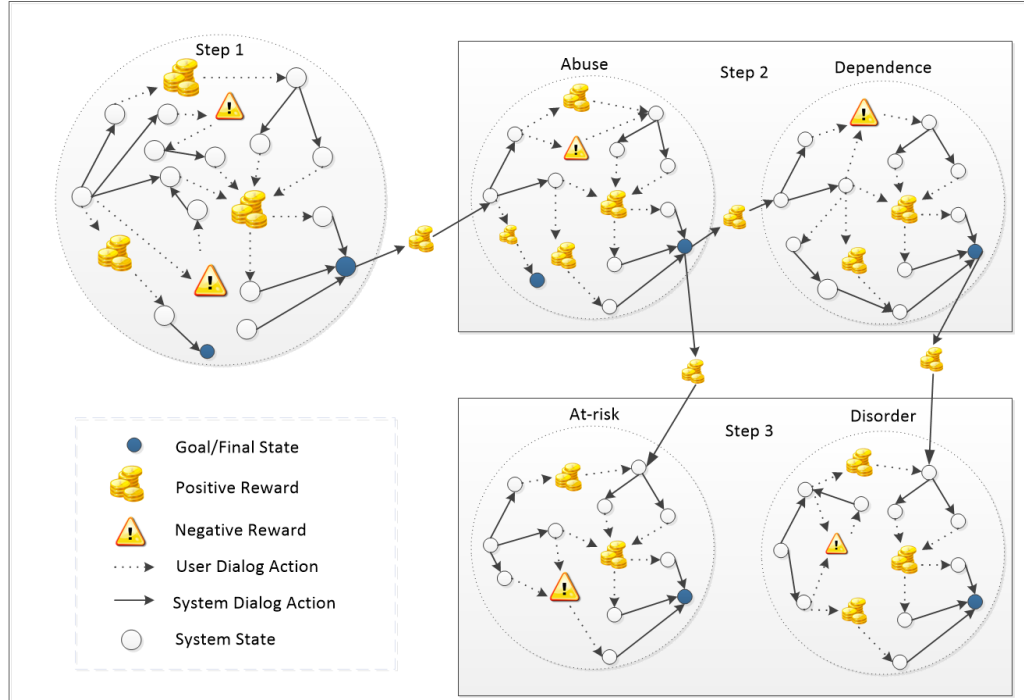


Figure 3.1: Representation Of World Model With MDPs

by user’s dialog actions in each dialog turn. For example, when the user speaks to the system, the speech recognizer *Confidence* level and *Value* attributes are updated based on the user’s dialog action. Our system aims to learn approximately optimal dialog strategies for the initiative style and the confirmation type selection.

3.4 Modeling World with Interconnected MDPs

To avoid the curse of dimensionality problem, I aimed at minimizing the number of system states used. Since the BI dialog requires many dialog turns between the system and a user, the number of available dialog strategies is very large, and can make learning optimal policies infeasible with limited training data. To alleviate this problem, I used separate MDPs for each phase.

I represent each step or phase of the BI with one MDP with local goals and reward functions. This approach divided the problem into 5 interconnected MDPs (shown in

Figure 3.1) but, in any interaction with the system, we use a maximum 4 MDPs, i.e. 1) Step 1; 2) Abuse; 3) Dependence; and 4) one MDP from Step 3 based on Abuse or Dependence problem. This approach also reduced the number of required state features for each step, thus reducing the number of states required.

Since there are two phases in *Step 2* (one for querying alcohol abuse and one for querying alcohol dependence), I represent Step 2 with two distinct MDPs (as shown in Figure 3.1), which greatly reduces the number of exploratory policies (because it reduces the number of state features) without compromising fine-grained distinctions between dialog strategies. Because the two phases are independent from each other, representing each phase with a separate MDP is appropriate. It also provides advantages in terms of learning dialogue strategies less amount of data.

There are two separate MDPs for representing the two different phases in *Step 3*. One is used for representing the model for “At-risk” drinkers who do not have alcohol use disorder problems (i.e. no abuse nor dependence). The second one is used to identify drinkers with alcohol use disorders.

In conclusion, the system is modeled with 5 MDPs. In each MDP, there are multiple terminal states. Some terminal states terminate the Step (such as the consent state), and some terminal states provide transparent transitions to the start state (or start state distribution) of another MDP (see Figure 3.1). At the same time, the agent receives a positive reward. The agent also receives immediate positive/negative rewards as showed in Figure 3.1. For details on immediate rewards, please see Section 3.6. With this approach, learning the optimal dialog strategy for an entire dialog is reduced to learning optimal dialog strategy for each of the MDPs.

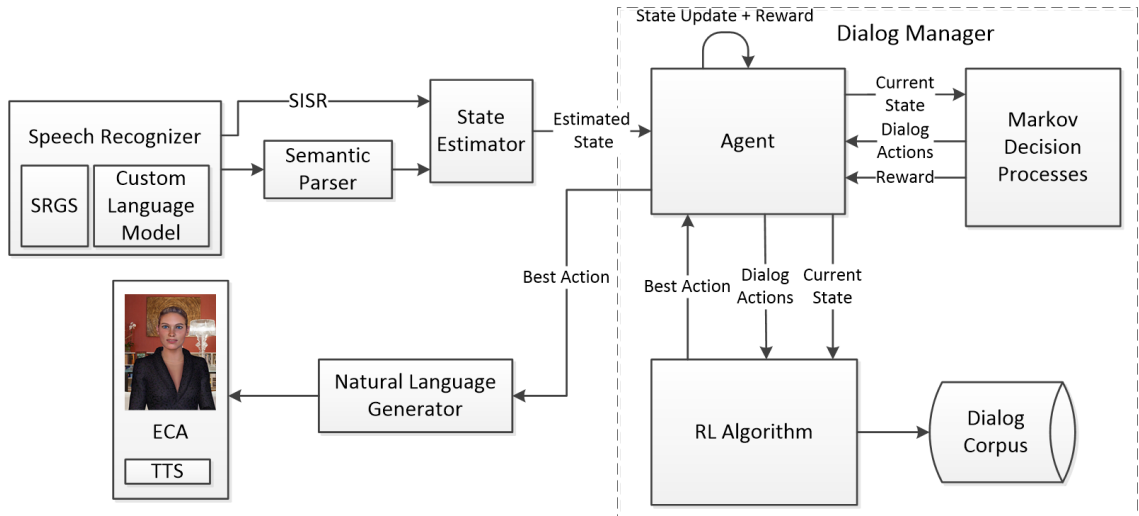


Figure 3.2: System Architecture and Dialogue Manager

3.5 Agent and Dialogue Strategy Learning

As shown in Figure 8.1, the *Agent* component of the system operates as an interface between other main components of the system. If the system asks a system initiative question, the *Speech Recognizer* component operates by using *Speech Recognizer Grammar Specification (SRGS) grammars*¹, and it outputs *Semantic Interpretation for Speech Recognition*² (*SISR*) tags. If the system uses non-restrictive grammar, it uses the *Semantic Parser* to parse the recognized speech. We use the Phoenix robust semantic parser [War91], which requires to write context-free grammar style recursive grammars to extract relevant information from the user utterances.

Therefore the *Agent* component receives *SISR* tags (i.e. when the type of system dialog action is system initiative or closed questions), or Phoenix parse results (i.e. when the type of system dialog action is user initiative or open questions) according to the initiative type, as semantic interpretations. The agent updates the system *Current State* and collects the *Reward* according to the reward function (see section

¹<http://www.w3.org/TR/speech-grammar/>

²<http://www.w3.org/TR/semantic-interpretation/>

3.6 for the reward function). It then queries the corresponding *Markov Decision Process* with the current state, and receives *Dialog Actions* and a *Reward* information for the current state, and there might not be any associated rewards.

A reward is received only if the *Current State* has an associated *Reward*. For example, the final state of each MDP has associated rewards. The agent sends the received *Dialog Actions* from the MDP and the *Current State* to the *RL algorithm*, and the RL algorithm selects the *Best Action* - *an action for which the agent received a maximum amount of reward in its prior experience* - based on the *Dialog Corpus* (see section 3.8) which is collected from real user interactions. The dialog corpus contains information about gained rewards at each step and accumulated rewards for a whole dialog session. The best action is the one that leads the agent to collect the maximum amount of reward. If the system is running in exploration/unooptimized mode, it selects dialog actions randomly among available actions in that state. Therefore, the best action selection does not happen in the unooptimized version which is usually used to collect training data (exploration mode).

The *Best Action* is passed to the *Natural Language Generator* component, which gives the final form of the system response and passes the text to the *Text-to-Speech (TTS)* engine. The embodied conversational agent *ECA* utters the response with lip synching. After each dialog turn, the *Dialog Corpus* is updated by the *Agent* with the old dialog state, action, the new dialog state and the reward information. Actually, the corpus contains more information about each turn but the RL algorithm uses reward signals to select the best dialog actions in each state.

At the inception of the project, I did not have any data for optimizing the system for our domain of discourse (the domain of alcohol use). So I first used the system with an algorithm which selects a dialog action randomly among the available ones.

Since I have mapped each state to sensible dialog actions, the system was able to deliver basic unoptimized functionality.

After having acquired the *Dialog Corpus* for the domain of alcohol abuse - which is itself a contribution as it can be reused - we used the RL algorithm to learn optimized dialog policies and select the best action according to available data (see section 5.2).

Based on each of our MDPs, the expected cumulative reward $Q(s, a)$ of taking *action* a from *state* s can be calculated in terms of Q-values of the next dialog states with the following equation [SB98];

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} P(s'|s, a) \max_{a'} Q^*(s', a'). \quad (3.1)$$

where $P(s'|s, a)$ is the transition model and $R(s, a)$ is the local reward function. The γ ($0 \leq \gamma \leq 1$) is the discount factor which is mainly used to indicate importance of sooner versus later rewards.

The Q-values in Equation 1 can be easily computed with a desired threshold using the Q-value version of the standard *Value Iteration algorithm* [SB98]. The algorithm updates iteratively the current value of $Q(s, a)$ based on the current Q-values, and it stops when the update yields a difference that is below the threshold. Once the Value Iteration algorithm is completed, approximately optimal dialog strategies can be selected by the system, which are essentially dialog actions with the maximum Q-values. The optimized dialog strategy must collect the maximum amount of rewards from future users.

The biggest challenge of this approach is in collecting enough human-machine dialog data to learn an accurate model. To avoid the data sparsity problem, we used minimal state representations and approximated the true state of the system during the interaction. Since the length of the dialog is long, a large amount of data is required to optimize the system. As we describe in Section 5.2, we run the systems

in two modes, training/exploration and testing. Training mode is for data collection, and in testing mode, the system uses optimized dialog strategies based on the data collected in training. Therefore, Equation 3.1 is used only for testing mode.

3.6 Reward Function Design

The reward function we use is designed based on the amount of information collected and the cost of collecting each piece of information. The agent gets a reward in each question: if the value is obtained in the first attempt with the *ASK* type of action, it gets +10 reward; if on the other hand the value is not obtained, the agent gets no reward. For each *Confirmation* action, if the obtained value is confirmed by the user, it gets +2, otherwise it gets -2. For each *ReAsk* action which could not result in obtaining the necessary information, the agent receives -3 reward, otherwise it receives +3 reward for the obtained value. If the obtained value is disapproved by the user, it deletes the previously gained reward. Therefore the agent gains a positive or negative reward for each question and dialog action. In addition to rewards gained per question, there are rewards in the MDPs which are associated with the final states. The system receives +15 reward if it is able to reach any of the final states in any MDP. For example, the successful completion of Step 1 gives the agent a +15 reward. In Figure 3.1, we depict the immediate rewards and the rewards that are received from the goal states for each MDPs.

I have used this approach to perform strategy learning for each question. Since the system tries to obtain one piece of information in each question, learning the approximately optimal actions in each question is useful.

3.7 Speech Recognition and Language Model

In the system, the operation mode of the speech recognizer³ is adapted according to the dialog manager's action selection. If the dialog manager asks **system initiative** questions to the user, the system uses *Speech Recognizer Grammar Specification (SRGS) grammars*. Even though I refer to system initiative questions as closed questions, the SRGS grammar does not restrict the user to answer with short answers such as yes/no or a number. It can still understand unrestricted speech. If the system operates in system initiative mode, the Phoenix parser is not used. Instead *Semantic Interpretation for Speech Recognition (SISR)* tags are used. We created a grammar by first authoring it in Augmented Backus-Naur Form (ABNF), and then we converted it to SRGS by using the NuEcho⁴ ABNF editor.

The system uses our custom dictation grammar while it operates in **user initiative** mode. In user initiative mode, we load two types of grammars in the in-process speech recognizer. One is the SRGS grammar which is prepared for the system initiative version of the current question. If the speech recognition result is based on dictation grammar, we use the Phoenix parser, otherwise we use SISR tags. Since the standard dictation language model is comprehensive, it does not work well in specialized domains. To address this problem, we created our own language model by using Windows Vista Dictation Resource Kit software. It is a tool which enables the creation of custom speech recognition dictation language models.

Language models help a speech recognizer decide upon the likelihood of a word sequence. Hence it is useful independently of the acoustics of the word sequences. A language model lets the recognizer make the right guess when two different sentences sound similar. For example, both of the following sentences sound similar: "Because of

³Microsoft Speech Recognizer

⁴<http://www.nuecho.com/en/>

alcohol, I had hard problems” and “Because of alcohol, I had heart problems”. With a language model on alcohol consumption, the recognizer knows that the first sentence is more likely to be what was said than the second one. Furthermore, a language model does not only give information about homonyms, it also gives statistical information about which word might appear after another, among other information. Therefore, if a language model consists of word sequences that are relevant in a specific context, it is very likely that it will operate better than a comprehensive language model for English.

To collect the data for the language model, we first collected data using the Mechanical Turk (MT) crowd sourcing website⁵ after obtaining Internal Review Board approval for the study. We asked MT participants the same questions that our system in full mode would ask (after being built from the process described above and after we have acquired the language model). In the instructions, we requested them to role play a person who is having alcohol problems. Our instructions were:

“Imagine that you are recently having drinking problems and that you are talking with a health professional face-to-face about your drinking problems. The health practitioner asks you the questions on this page. Please answer as naturally as possible.”

Because alcohol usage is a very common and universal social problem that everyone understands, MT users’ answers were relevant. One can note that we would not necessarily have collected meaningful answers had we asked MT users, for example, to imagine having some complex disorder such as schizophrenia, because most people do not know what behaviors are associated with this condition. Consuming alcohol in different quantities however, is an experience that many people can relate to, and therefore the answers that we collected were very relevant.

⁵<https://www.mturk.com>

Participants answered the 18 questions. We created the language model from the responses of 447 MT workers. We preprocessed it (corrected spelling and grammar problems) before creating the language model. We improved the language model by adding sentences generated based on our SRGS grammars, and used this language model in our experiments. In the model, there are 7,599 utterances, the average length of an utterance is 11.82 words, there are 100,679 word tokens, and 5,423 distinct words.

We used our custom language model in our evaluation (see section 5.2). We collected the training data from real user dialogs (described in Section 3.8) which includes sound files. We ran the speech recognizer on the collected sound files and compared recognitions based on the two language models. We performed quantitative analysis to compare the Microsoft standard dictation language model with our custom language model. We found that when we use our custom language model, the word error rate is approximately 17% lower than the Microsoft standard dictation language model.

3.8 Dialog Corpus

I created a very richly annotated XML-based dialog corpus from the test dialogs, whose size will continue to grow as we collect more data. The corpus is organized turn by turn. Each turn element contains: step and state information, question asked by the system, initiative type, best speech recognition, grammar type, semantic value or result of the Phoenix parser, N-best recognitions with confidence score, reward gained from the question, cumulative reward and sound files. Each XML log file contains sequences of dialog turns for one dialog session.

CHAPTER 4

Spoken Dialogue System Based on POMDP and Health Screening Dialogues

Latest progress in speech recognition technology, together with advances in the field of conversational intelligent virtual agents (IVA), have created new possibilities to develop a variety of useful applications to address contemporary health-care challenges. Because current automatic speech recognizers (ASR) are still regarded as a noisy input channel, they need to be backed up with a mechanism to operate against noisy recognitions. As discussed earlier chapters, in the spoken dialog systems (SDS) area, latest research has mostly concentrated on addressing this problem by employing stochastic and data-driven dialog management (DM) methodologies, namely reinforcement learning based approaches [YGTW13, FL09]. Whereas there is growing interest to develop speech-enabled conversational applications among IVA researchers [STA⁺10, THSh⁺11], speech has not been a widely used input channel. Speech-enabled interfaces has great potential to address health-care problems by increasing accessibility of health-care services for patients. Speech-enabled interfaces, however, face a variety of challenges, such as operating in noisy environments, expected low cooperativeness of users, accuracy and precision requirements for the received information. Although research in spoken dialog systems area has a potential to foster development of health dialog systems, the approaches that are used in SDS area for dialog management, are not widely used in the development of health dialog systems, except few examples [LGRT13, YLR13].

The work in this section aims to bridge the gap between SDS and IVA research and use findings from the SDS community for DM in the applications of IVAs in the health domain. More specifically, the concentration in this section is twofold. First, I created a dialog manager design tool which enables the creation of stochastic data-driven

dialog managers for structured health dialog interviews. The tool can be used to create spoken dialog systems for initial screening of patients, conducting brief health interventions and information-providing applications. It employs the techniques that are used in latest SDS research which provide resiliency while operating under noisy ASR outputs. Second, I developed an SDS integrated with a multi-modal embodied conversational agent for delivering an alcohol screening and brief intervention for youth.

The work in this chapter also improves the ideas presented in Chapter 3 by using partial observable Markov Decision Processes instead of fully observable MDPs. In this section, I still work on alcohol domain but for youth instead of adults to demonstrate of applicability of the tool to similarly structured health interviews.

4.1 Underage Drinking Problems and Computer-based Interventions

Alcohol known as the most commonly used and abused substance, is the cause of 4,300 annual underage deaths [fDCC⁺08]. Although drinking under the age of 21 is not legal, 11% of all alcohol consumed in the United States is drunk by people aged 12 to 20 years [fRE05] and more than 90% of it is consumed in the form of binge drinks [fRE05]. In 2010, underage alcohol consumption caused 189,000 emergency room visits because of injuries and other conditions linked to alcohol use [Abu13].

The magnitude of social costs of underage drinking is very high. It has been evidenced by studies that computer-delivered interventions have positive effect on reducing alcohol consumption and binge drinking among underage drinkers [WMM12, CSSE⁺09]. Delivering interventions using conversational interface with ECAs has been shown to increase engagement and intention to use behavior change systems [LAYR13], and have positive impact on the acceptance of computer interventions.

4.2 Alcohol Screening And Brief Intervention For Youth

In this section, I introduce the domain and the content of the system. I use the pocket guide from National Institute on Alcohol Abuse and Alcoholism (NIAAA) for alcohol screening and brief interventions for youth [oHS11], which is publicly available online. The system strictly follows the guidelines specified by the pocket guide¹. For the questions that the system ask, I use the content in the full version.

The initial screen and brief intervention has 3 steps: **1)** Step 1: Ask the two screening questions; **2)** Step 2: *a)* Guide Patient, *b)* Assess Risk **3)** Advise and Assist.

In *Step 1: Ask the Two Screening Questions*, the system asks the age or the type of school the person is pursuing, if friends of the user is drinking or not, and if user is drinking or not. According to the age or the school type. The system stores the inputs that it receives for the first 3 questions in the dialog history. If the patient does not drink, the system goes to *Step 2a: Guide Patient*. If the friends of the patient drink, the system takes different path. Without depending on friends' drinking status, it reinforces healthy choices, elicits and affirms reasons to stay alcohol free and educates the patient about alcohol consumption. For a patient who does not drink and with no drinking friends, it ends the conversation by asking for re-screening the next year again. For a patient who has drinking friends, it explores the patients' plans to stay alcohol free when friends drink, advise and offer to re-screen next year. If the patient drinks, it goes to *Step 2b: Assess Risk*.

For a patient who drinks, the system transits to *Step 2b: Assess Risk*. For a patient who has high risk, the system conducts brief motivational interviewing. Brief motivational interviewing roughly queries patient about drinking patterns, problems

¹The full NIAAA guide is 40 pages, and the pocket guide is a lot shorter summarized version of it.

experienced or risks taken. It also asks if parents know about the patient’s drinking issues and refers the patient to treatment. For a patient with moderate drinking, it gives brief advice, if patients does not have alcohol problem. If a patient has an alcohol problem, it conducts brief motivational interviewing as in the case of high risk, asks if parents know and arranges follow-up. The system provides brief advice, and reinforces healthy decisions. The details of the intervention can be accessed in NIAAA website².

4.3 System Architecture Overview

As can be seen in Fig. 4.1, the spoken dialog system contains *Speech Recognizer*, a *text-to-speech (TTS)* engine as user interface combined with a *3-dimensional animated ECA*. The speech recognizer operates in two modes, system initiative (i.e. restricted grammar) and user initiative (i.e. non-restrictive grammar). If the system asks a system initiative question, the *Speech Recognizer* component operates by using *Speech Recognizer Grammar Specification*³ (*SRGS*) grammars, and it outputs *Semantic Interpretation for Speech Recognition*⁴ (*SISR*) tags. For the *Custom Language Model* I used a crowd sourcing service⁵. A statistical language model assigns a probability to a sequence of n words by means of a probability distribution. Language models help a speech recognizer figure out how likely a word sequence is, independent of the acoustics. This lets the recognizer make the right guess when two different sentences sound the same. I asked the questions the system asks to Mechanical Turk workers. I collected answers from 1,075 participants to create the language model. I use the Phoenix *Semantic Parser* [War91] for extracting relevant information from

²<http://www.niaaa.nih.gov/Publications/EducationTrainingMaterials>

³<http://www.w3.org/TR/speech-grammar/>

⁴<http://www.w3.org/TR/semantic-interpretation/>

⁵<https://www.mturk.com>

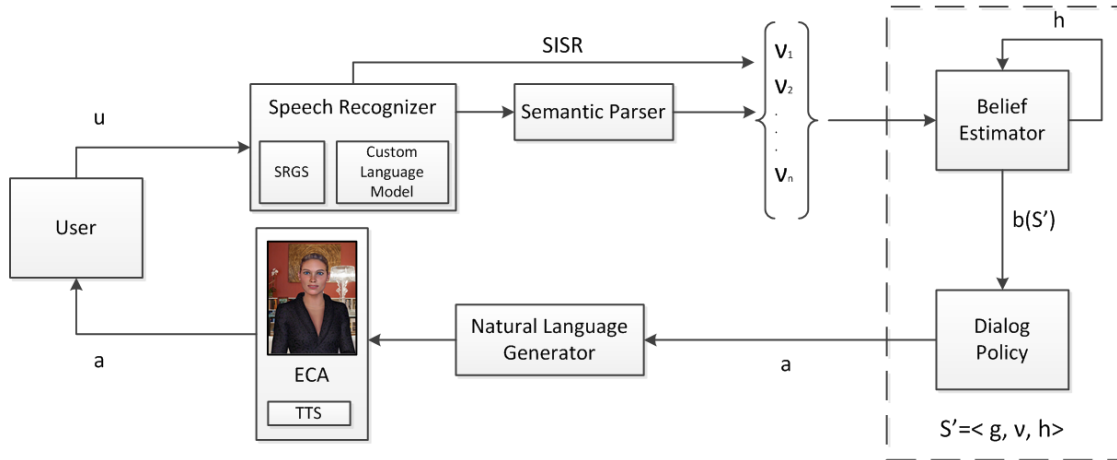


Figure 4.1: Architecture of the multimodal spoken dialog system: u and a denote user and system dialog actions, v_n denotes estimated user dialog action, h denotes dialog history, b is systems belief for the estimated state S' .

the Speech Recognizer outputs. Belief estimator calculates the estimated dialog state based on n-best speech recognitions and the dialog policy component decides for the next system dialog action. I describe the methodology that I use for the dialog management (i.e. belief estimator and dialog policy components) in Section 4.4 in detail. The *Natural Language Generator* (NLG) module contains templates for system dialog actions. According to dialog action selection, the NLG module gives the final form to the dialog action which is the text to be spoken by the ECA. Microsoft Speech API⁶ compliant TTS engine generates phonemes for the ECA system. The ECA system speaks to the user with lip synchronization and some simple non-verbal behaviors (e.g. head movements).

4.4 Approach

As I have indicated before, health interview dialogs are different than dialogs found in the tourist information domain which is the most widely domain studied by SDS

⁶[http://msdn.microsoft.com/en-us/library/ee125663\(v=vs.85\).aspx](http://msdn.microsoft.com/en-us/library/ee125663(v=vs.85).aspx)

researchers. The main goal of health interview dialogs is to collect initial screening information, educate patient, increase awareness and, if needed, refer the patient to a treatment. There is usually published plan of standardized health interviews (e.g. [HHEA⁺10, SRC⁺05, NIA07a]) by national or international health institutions. So the system has to conduct the conversation according to that plan. The system usually needs to ask one question at a time and in a specific order, while the flow of the dialog adapts according to the received answers. In other words, the inputs received by the system influence the future steps of the conversation (i.e. sequential or branching). Unlike tourist information systems that do not involve different dialog execution plans such as branching dialog plans or conditional steps that might be followed based on answer of single or multiple questions, the system does.

Standardized health interviews have a common pattern and structure. The plan of the interview might contain 3 patterns; 1) *sequential independent questions* (where the future questions do not change according to current answer); 2) *branching questions* (each different answer causes the system to ask a different question); 3) *composite questions* (asking a composite question depends on answer of multiple prior questions). A standardized health interview might contain one or multiple patterns together.

While designing a dialog manager for standardized health interviews, the system designer needs to consider all three patterns based on the standardized assessment/intervention guide. The purpose of asking each question might aim at receiving one (i.e. closed question) or more (i.e. open-ended questions or mixed-initiative questions) pieces of information. However, at the very least, each question aims at receiving a single very explicit piece of information (i.e. closed questions) from the user. This requirement affects the operation mode of the Speech Recognizer.

For the open-questions, the Speech Recognizer uses unrestricted recognition grammar, whereas for closed questions, it use restricted grammars (e.g. SRGS). Therefore,

the system actions changes according to the dialog strategy. For example, a system might use an open question but if it can not recognize (e.g. because of low confidence scores) and/or extract (i.e. this operation is performed by the semantic parser module on ASR outputs) the expected information, it might re-ask the question in a closed-style to get the minimum required information. Re-asks and implicit or explicit confirmations are widely used system actions in information-seeking dialog systems. Therefore, the type of system actions that are used to get each piece of information, are usually all the same.

4.4.1 Belief Monitoring and Partially Observable Markov Decision Processes

Each system has to track a dialog state which usually contains some important dialog state attributes such as: ASR confidence level, grammar type, information about whether the received answer is confirmed or not, number of re-asks. Each state is mapped to a dialog action that is called *Dialog Policy*. To optimize the system, a reward function is designed. I can formalize the defined mechanism with the Markov decision processes (MDP) framework which described earlier in Chapter 8.3. I briefly described below again.

MDP framework can be characterized by a tuple (S, A, T, R) , where: S is a finite set of states, A is a finite set of actions, T is a state-transition function such that $T(s', a, s) = p(s' | s, a)$ which describes the probability of performing action \mathbf{a} in state \mathbf{s}' will lead to state \mathbf{s} , $R(s, s')$ is a local reward function, and the objective of the SDS is to maximize the gained reward.

MDP assumes that the entire state space is fully observable. However, it is partially observable in SDSs because of imperfect ASR outputs. An SDS with Partially

Observable Markov Decision Processes (POMDP) model attempts to address the partially observable nature of SDS state spaces [WY07]. According to SDS-POMDP model, at each dialog turn, a user has a goal g in mind (e.g. answering with the number of days in week s/he drinks, or whether the alcohol causes any health problems). The system takes a dialog action a (e.g. how many days in a week do you drink alcoholic beverage?) and user replies with action u ("I usually drink on weekends"). The speech recognizer outputs the N-best list of recognitions $\nu = \{\nu_1, \nu_2, \dots, \nu_n\}$ with the estimated confidence scores indicating the likelihood of each recognition being correct $P(u|\nu)$ while processing the audio. A history variable h keeps track of the relevant dialog history (e.g. receipt of each piece of information, confirmation status of each piece of information). Because ASR is a noisy sensor, g , u , h are not fully observable by the system. Instead, the system maintains a distribution over these values b . Given some existing distribution $b(g, h)$, and observations a and ν' , an updated distribution $b'(g', h')$ can be computed [WY07, Wil10]:

$$b'(g', h') = k \sum_{\nu'} P(u'|\nu') \sum_h P(u'|g', h, a) P(h'|g', u', h, a) \sum_g P(g'|a, g) b(g, h) \quad (4.1)$$

where $P(u'|g', h, a)$ gives how likely user actions are; $P(h'|g', u', h, a)$ gives how the dialog history evolves; and $P(g'|a, g)$ gives how the user's goal may change. k is a normalizing factor.

POMDPs grow exponentially with the number of possible user goals, and it is not possible to calculate this update in real time. This means that POMDP usually suffer from scalability issues [WY07, YGK⁺10b]. To overcome this problem, a distribution over the set of partitions of user goals $\{p_1, p_2, \dots, p_n\}$ is maintained, each partition p_n indicates a collection of user goals, and each user goal can be belong to exactly one partition. The belief in a partition is the sum of the dialog states it contains.

It is assumed that the user’s goal is fixed during the interaction, and that error-prone ASR confusions between recognitions not on the ASR N-best list are uniform. These two assumptions allows to compute [YGK⁺10b]:

$$b'(g', h') = k \sum_{\nu'} P(u'|\nu') \sum_{h \in p'} P(u'|p', u', h, a) P(p'|p) b(g, h) \quad (4.2)$$

where $P(p'|p)$ shows the fraction of belief in p which p' would have if p were split into p' and $p - p'$: $P(p'|p) = b_0(p')/b_0(p)$ and $P(p - p'|p) = b_0(p - p')/b_0(p)$, where $b_0(p)$ is the prior probability of a partition p [YGK⁺10b].

The partitioning is performed in the following way: first each recognition in the N-best list is compared to each existing partition; if user action can split the partition, the partition is divided. Then the belief in each partition (and dialog histories) is updated using Equation (2). To avoid exponential growth of the number of partitions, low confidence partitions are combined by summing up their beliefs. This approach usually allows to take into account 2-3 N-best recognitions [YGK⁺10b]. This problem is addressed by applying incremental partition recombination for tracking dialog state by using a larger number of N-best recognitions [Wil10]. I use the incremental partition combination approach in dialog state tracking [Wil10].

4.4.2 Representing Patterns in Health Interviews As Programmatic Objects

The goal of my approach is to create a tool to design custom dialog managers which employs POMDPs as an underlying mechanism. As a result, a dialog designer can just specify the content-related information (e.g. question to be asked, an information that need to be provided) and connections between the each question in the interview. To facilitate the process, we created an API to design dialog managers.

Each question in the intervention represented as an object which encapsulates the dialog policies and the transition information. In other words, each question object consist of a POMDP with transition information to successor step. Since the most of the dialog actions have similar purposes such as asking a question, confirmations, and re-asking a question, it is possible create parameterized patterns that are encapsulated in programmatic objects. Basically, each question object contains all the underlying basic functionality for each piece of information which can be customized. The questions objects can be considered as nodes of a graph, the transitions between questions can be considered as edges that requires a key value to transit from one node to another. The key value is a piece of information that the system tries to get from a user in a particular question. I described details of the design later in this section, but Fig. A.4 is an example of the described structure.

In my design, I considered the 3 dialog patterns (i.e. linear/sequential, branching, and composite) for health interviews that I covered in Section 4.4. Each question (i.e. system dialog action which might be question or sometimes a piece of information) is represented as a question programmatic object. Each question object contains question specific information; **1)** question text which might contain several versions according to initiative type; **2)** type of node (i.e. root or not root) **3)** type of successor nodes (i.e. single or multiple); **4)** semantic key types (i.e. any, specific, composite): *any* is used for linear/sequential transitions where the next transition does not depend on any semantic key, *specific* type is used to indicate that the transition branching is based on simple semantic keys that are extracted from user actions, *composite* keys indicate that transition depend on answer of multiple prior questions; **5)** semantic keys; **6)** dialog actions such as ask, confirm, re-ask; **7)** *edges* which are connectors between other question objects. Each question object contains the default dialog actions (confirmations, re-asks) except the specific question (i.e. because style

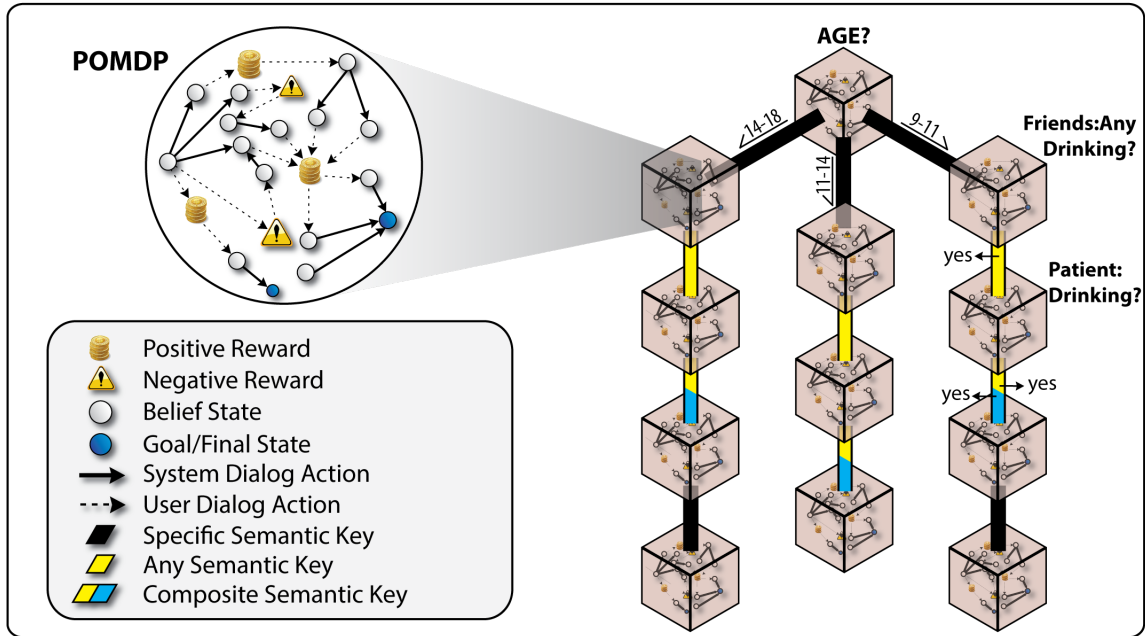


Figure 4.2: Demonstrates sample dialog structure that could be created with the dialog design tool which includes three different dialog patterns that do exist in the alcohol intervention dialog for youth. Cubes represents question objects, edges between cubes represents semantic keys to transit from one question to another, circled area shows underlying POMDP mechanism for each question object.

to request information can be generalized for other actions but not for the actual question) however each action can be modified by the user. In addition each question object contains underlying POMDP mechanism that is discussed in section 4.4. The questions are connected to each other by directed edges (see Fig. A.4) which have a key that represents the required input to transit to the next question (i.e. the required input could be specific, composite or any).

A dialog designer should instantiate a question object with at least 4 parameters **1)** the question text, **2)** the semantic keys which are used to create edges from current node to successor nodes, **3)** the prior node which indicates which is the prior of the current node, and **4)** the semantic-key-to-connect indicates which edge of the prior node it should connect. If it is desired, the default dialog actions included in a question object also can be modified.

The dialog agent makes transitions between the questions using received semantic keys and, in each question, the internal state transitions depend on POMDP mechanics. I depicted a sample dialog plan structure in Figure A.4. An agent has knowledge about which question it is working on but it only has belief about what state it could be in POMDP for that question.

As a concrete example of a dialog design, I use NIAAA pocket guide which is described in Section 4.2. The first question that the system needs to ask is the age of or the type of school where the person studies. Since it has 3 possible answers and the required number semantic key for transition is one, and each one will transit to different question, the dialog pattern type is *branching* and the semantic key type is *specific*. The system developer should also specify the type of node, if it is root node or not (the first question is always root node). When the system developer (SD) specifies all of the required parameters. The system creates a question with its edges to connect to successor questions.

For the elementary school students, the SD needs to create a new question object to ask if any of the student's friend are drinking alcoholic beverages. This question has a prior question which is the root question. The type of answer will not change the next question that will be asked by the system, in other words there is no branching. Therefore, the type of successor is *single*. Semantic key type is *any*, because the type of answer to transit successor does not matter.

The next question is *branching* (if patient drinks or not), however, according to the answers that are received to the current and previous questions and the system transits to *step 2: guide patient*. Branching is done based on the *composite* key (patient drinks or not, patient's friends drink or not). Therefore, the third question successor type is *multiple* and semantic key type is *composite* (i.e. I showed the composite key but can't show the branching in the figure because of the space limitation).

Creating question objects which encapsulates POMDP mechanism is as easy as specifying some content-related parameters. Also for each dialog action, the system developer needs to create SRGS and Phoenix grammars. While the system works on a question, it tracks the dialog state seamlessly using the stochastic approach that I discuss.

4.5 Discussion

I created dialog manager design tool for creating dialog managers for delivery of standardized health interviews. It will increase the accessibility of state of the art dialog management approaches to non-experts. Also, I adapted the latest methodologies that are used by the SDS community to health dialogs. I discussed my implementation for alcohol screening and brief intervention for underage drinkers. As future work, I plan to test the system with real users for both objective and subjective measures. For the programming tool, I plan to create functionality to handle additional dialog patterns.

CHAPTER 5

Experiments with Optimized and Random Dialog Policies

In this section, I discuss my experiment that is conducted to evaluate spoken dialog system with real users. The evaluation concentrates on two aspects of the system, objective and subjective aspects.

5.1 Sample Population

University students represents very appropriate sample for target population for brief interventions. The latest report of NIAAA on college drinking indicated that alcohol problems are very prevalent among college students [NIA07b] and 19% of college students (18-24 ages) met criteria for alcohol abuse or dependence¹. The use of brief interventions with college students to educate students about drinking and their increase awareness is very common [NIA07b]. As a result of many studies, the NIAAA report on college drinking emphasized that "increased alcohol screening and brief interventions are feasible and appropriate for identifying and addressing harmful drinking among college students".

In addition, using computer and web-delivered interventions is very well studied in college settings [WHCM05, WN05, SPF⁺07]. For example, Saitz et al. [SPF⁺07] tested the feasibility of providing online alcohol screening and brief intervention to more than one-half of an entire freshman class. The students were contacted through e-mail and invited to take the brief intervention. The researchers found that, in general, unhealthy alcohol use - ranging from risky drinking to alcohol abuse and dependence - decreased following the intervention. Hence, although we are not assessing the impact of the system on health/drinking outcomes (which would require

¹From the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSMIV), American Psychiatric Association.

a randomized clinical trial outside the scope of this study), our target population is very appropriate for participating in brief interventions.

For the evaluation of the system, 89 subjects were recruited from volunteer university students through fliers and emails. From 89 participants, 62 of them were males and 27 of them were females; 51 of them were native speakers and 38 of them were non-native speakers, which realistically represents the diversity of the population in the Miami, Florida area.

Participants sat in front of a PC computer running the systems (some the training system and some the testing system as described below), and responded in English to the questions asked by the embodied conversational agent shown in Figure 2.3. The computer was equipped with a USB sound card and a Sennheiser ME 3-ew microphone.

It is important to note that we did *not* perform any user training nor speaker adaptation for speech recognition.

After obtaining an oral consent approved by the University Internal Review Board, we gave the following instructions to each subject before using the system for both experiments:

- You will be asked questions about your drinking behavior with an avatar/virtual character. You may or may not have any alcohol related problems, but we just want you to role-play a person who is having drinking-related problems and give relevant answers to each question.
- Try to speak clearly and loudly enough.
- Wait until the avatar stops speaking before you answer.

5.2 Evaluation Results

In the *first phase* of the study, for the first 52 subjects, the system operated in training/exploration mode and selected random dialog actions from the available ones in each state (see section 3.5 for discussion). In the second phase, the remaining 37 subject used the system in testing mode. Since, we mapped each state to sensible dialog actions, the system could deliver basic, but expectedly unoptimized functionality. The goal of the first phase was to collect training data to optimize the system for initiative and confirmation type selection.

In the *second phase* of the experiment, the users used the optimized system. Even though the number of subjects is not very large to compute the optimal dialog strategies, it was sufficient to compute approximately optimal dialog strategies. I observe the positive effects of optimization while testing the optimized system.

5.2.1 Task completion evaluation

In Table 5.1, we present the results of our *task completion evaluation*: Column 1 “Evaluation Measure” is the type of evaluation; Column 2 “Training” is the mean of task completion measure obtained for the training system; Column 3 “Testing” is the mean of task completion for the optimized system; Column 4 “ Δ ” shows the difference between testing and training averages; and Column 5 “p-value” is the statistical significance value obtained using the standard two-sample t-test over subject means.

I show the average values of binary task completion across 52 training dialogs and 37 testing dialogs. At the end of the each interaction, I asked questions to each subject. One of them was “Did you complete the intervention?”. If they completed the intervention, the binary completion value was +1, otherwise it was -1. The task

Table 5.1: Task Completion Rate: Training versus Testing

Evaluation Measure	Training	Testing	Δ	p-value
Self-Report Completion	0.1538	0.5675	0.4137	0.0402
Real Completion	0.03846	0.4594	0.42094	0.0434
Step1: Assessment	0.3461	0.7297	0.3836	0.0371
Step2 Abuse	0.3076	0.6216	0.3139	0.1058
Step 2 Dependence	0.1923	0.6216	0.4293	0.0300

completion reported (and perceived) by the subjects is referred to in Table 5.1 as *Self-Report Completion*.

The additional *Real Task Completion* measure is defined because perceived task completion and real task completion are different. Real task completion indicates whether or not the system obtained all the answers for each question it asked. The perceived (self-report) task completion is different because, if the system can not obtain the answer in three attempts, it skips that question without having an answer and the user is not aware of it.

Three other task completion metrics show the real task completion for each step. The training and testing blocks show averages of binary task completion for each individual version of the system. Since the difference between “real completion” and completion rates for *Step 3* is negligible, we do not report it.

Each row shows a different task completion information and compares the two versions of the system. The first row is the *Self-Report Task Completion* (perceived) for the whole intervention. The difference between the two versions is statistically significant ($p = 0.0402 < 0.05$)². As mentioned above, the perceived task completion refers to when the subject could complete the intervention, even though there may

²Conventionally, a p-value less than 0.05 is considered statistically significant, a p-value less than 0.10 is considered indication of a statistical trend.

exist some questions which the system could not obtain answers to, but the user was not aware of it.

The second row shows the *Real Task Completion*, which means that the system did obtain an answer for each question asked. The mean values are lower than self-report completion because the system was able to complete sessions by skipping questions. For example, according to the NIAAA guide for brief interventions which we followed (see details above) [NIA06], it is enough to obtain 1 abuse indicator with the 4 questions which query alcohol abuse. If the system could not obtain an answer to the first three question but obtained an answer to the forth one, the user could still complete the session but from the system’s perspective, there are questions which it could not obtain answers to. The difference between the training and the testing system for real task completion is statistically significant ($p = 0.0434 < 0.05$).

The difference in task completion rate for the *Step 1: Assessment* is statistically significant ($p = 0.0371 < 0.05$) for the training and testing versions. Step 1 contains five questions, and since the dialog length is short, a higher task completion rate is expected for both of the versions.

The difference in task completion rate for the *Step 2: Abuse* is not statistically significant ($p = 0.1058 > 0.05$). This is because of the length of the this step. However, as mentioned in the NIAA guide for brief interventions [NIA06], it is sufficient to find a 1 abuse indicator to pass to the “Step 2 Dependence” step.

The difference in task completion rate for *Step 2: Dependence* is statistically significant ($p = 0.0300 < 0.05$). This step is long and the system needs to identify three indicators by using 7 questions. The task completion rates for each sub-steps of *Steps 3: Advise* converge to real task completion rate because it is the end of the intervention. Since the difference between real completion and completion rates for *Step 3* is negligible, we did not report it.

The task completion rate in the training dialogs is 58%, and for the optimized system it is 77%, an improvement of task completion rate of 19%. Although the results we obtained are statistically significant for most of our task-completion criteria, for data hungry reinforcement learning algorithms with a large number of system states, a larger number of subjects will allow us to draw conclusions about the optimality of the learned policies. However, as shown in Figure 5.1, we compared Q-values for each episode. An episode can be defined as completing one question and passing to the next question. Completion of a question does not mean that the system obtained the information it was trying to get. As discussed earlier, it is possible for the system to transit to the next question without having obtained the information, and in that case, the system receives negative reward. We described the details of the reward function in Section 3.6. We show the improvement of Q-values for each episode in Figure 5.1. We have 21 episodes because we have 18 questions, plus transitions between MDPs. As shown in Figure 5.1, the optimized policy performed better, even though it is not optimal. We have to note that *optimal policy* represents the highest reward that the system can achieve, whereas the *random policy* and the *optimized policy* represent the average score that the system collected in training and testing operation modes, respectively.

5.2.2 Dialogue Evaluation

In addition to task metrics, we looked at *Dialog Metrics* to measure the number of turns for successful completions, and the number of words per turn.

The average length of a dialog is 31.9 turns, the shortest completed dialog is 24 turns and the longest one is 43 turns. The length of the dialog is significantly larger than similar RL-based systems [SLKW02, YGK⁺10a, FL09]. For Step 1, Step 2 abuse, Step 2 Dependence and Step 3, the average length of the dialog are respectively: 9.6,

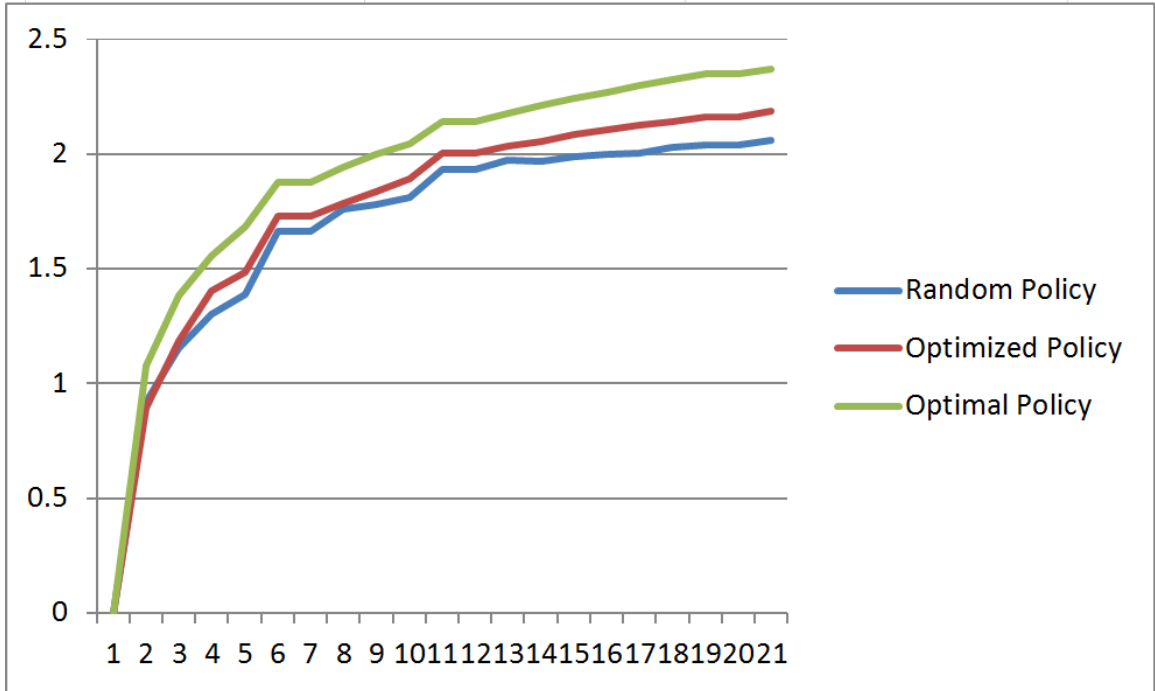


Figure 5.1: Q-values for each episode, x axis shows episode and Y axis shows log-scale Q-values.

4.8 and 13.4, and 4.1. The average number of words used or recognized in each turn is 3.3.

5.2.3 Subjective Evaluation

After the subjects completed the intervention, the subjects answered a survey aimed at evaluating the user’s experience with the system. The survey has two parts, the first part has 4 yes/no questions and the second part is a 34-item questionnaire about the subject’s assessment and experience with the system.

In the first part, we asked questions about reuse “Would you use the system in future?”, and ease of use “Is the system easy to use and is it easy to understand how to use the system?”, and “Did the system understood what you said” and “Did you know what to say to the system in each turn”. Since these 4 questions are not

directly related with dialog strategies and we want to see the complete picture, we did not compare test and training systems.

The evaluation of the subjective aspects shown in Figure 5.2 demonstrates that acceptability of the system by users is very high in terms of *Ease of Use* (81 Yes versus 8 No) and *Intention to Reuse* (63 Yes versus 26 No) the system. The *What to say to system* shows that sometimes users do not know how to answer the system questions. We believe that the reason can be that when the system is in user initiative mode (open questions), the subjects may not be sure to what extent they should provide details.

The *System understood* criteria shows that most of the users think that the system understood what they said. We postulate that this is achieved with our ample use of confirmation questions that the system utters when not sure.

In the second part of the subjective assessment, we used a 34-item questionnaire named *Subjective Assessment of System Speech Interfaces (SASSI)* [HG00]. It is a widely used evaluation questionnaire in the SDS community. The subjects answered a randomized list of SASSI questionnaire on a 7-point Likert scale. The SASSI questionnaire queries 6 aspects of the user's assessment and experience with the system. These aspects are *Accuracy*, *Likeability*, *Cognitive Demand*, *Annoyance*, *Habitability*, and *Speed* of the system.

The items in *Accuracy* are related to whether the system recognizes user's input correctly and does what the user expects. The items in *Likeability* are related to whether the system recognizes user's input correctly and does what the user expects. The items in *Likeability* include statements about the opinion and feelings of the user about the system. *Cognitive Demand* summarizes the level of effort needed to use the system and the user's feelings arising from this effort. The *Annoyance* includes statements such as "the interaction with the system is repetitive/boring/irritating".

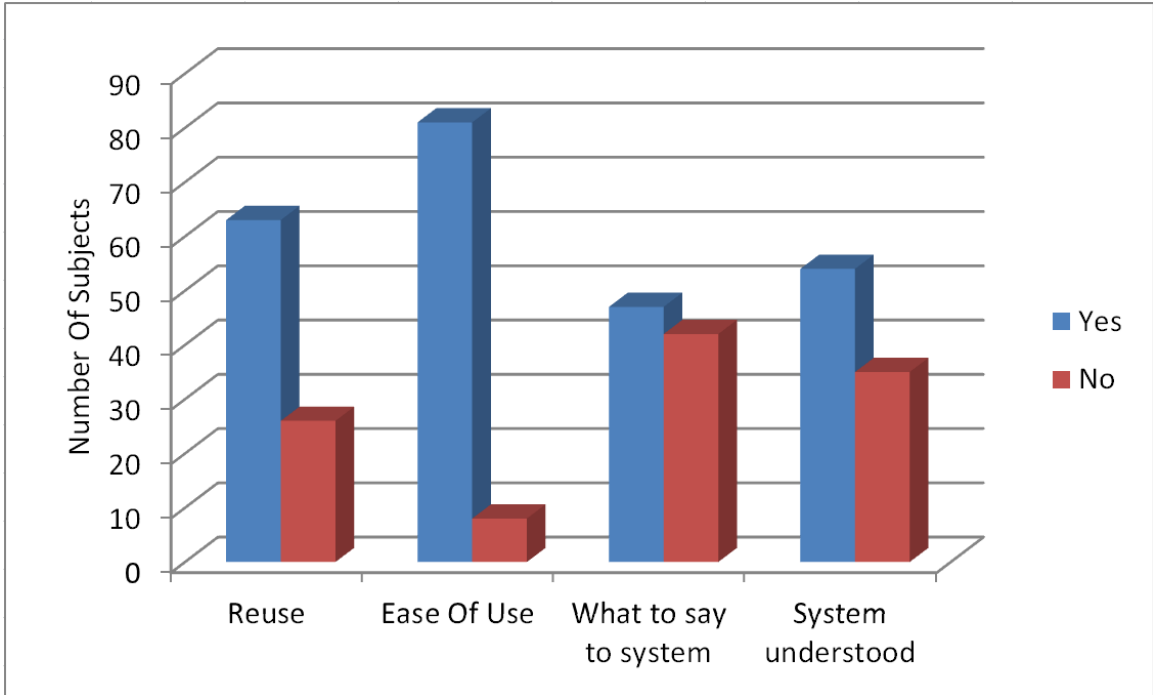


Figure 5.2: Subjective Evaluation

Habitability contains statements related to whether the user knows what to say and knows what the system is doing. The *Speed* contains only 2 items related to the speed of the system.

We compared two versions of the system (training and testing) for the SASSI evaluation. As discussed earlier, **52** subjects used the training system and **37** subjects used the testing version of the system. We show the results in Figure 5.3. In the 7-point Likert scale, 1 is the lowest negative score (strongly disagree), 4 is neutral score (neither agree nor disagree) and 7 is the highest score (strongly agree).

We actually compared two versions of the system but our goal was also to assess the overall performance of the system for speed and habitability categories, because both versions of the system do not have any difference in terms of features which are assessed by speed and habitability measures. To be consistent, we compared habitability and speed measures, as we did for other subjective measures. The results

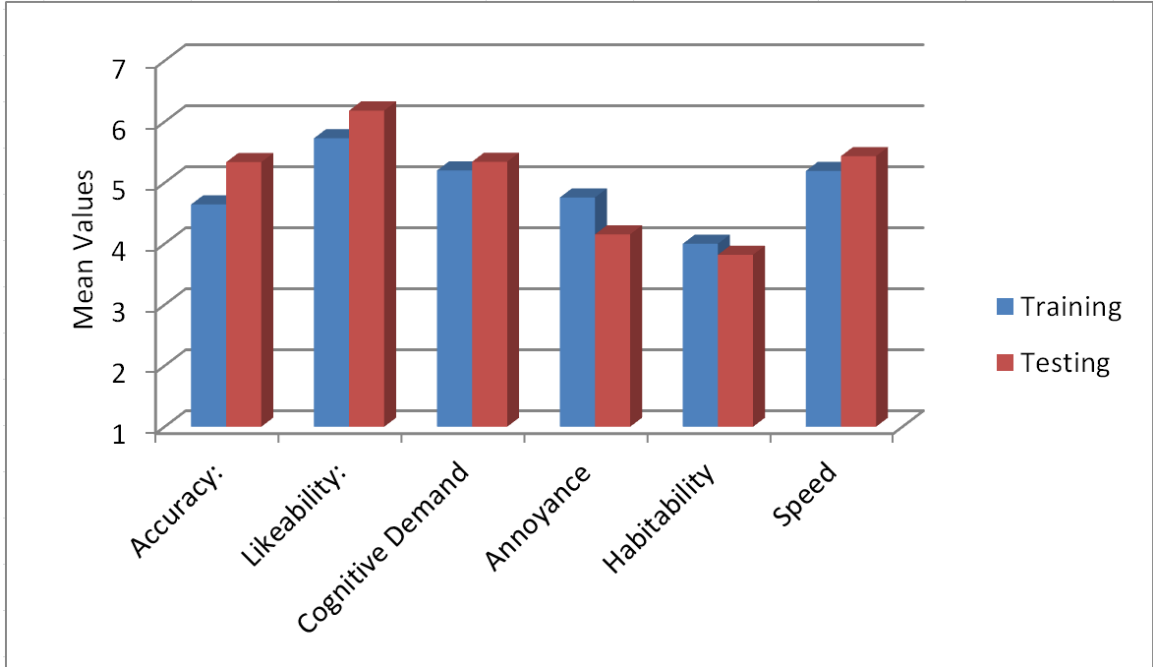


Figure 5.3: Assessment - Negative (1) Neutral (4) Positive (7)

for habitability and speed correlate our viewpoint, because the mean values are very close, as showed in Figure 5.3.

In Table 5.2, we show mean values for each evaluation category for both versions of the system, the difference between mean values, and p-values. We obtained p-values by performing the standard two-sample t-test. Column 1 is the type of evaluation; Column 2 is the mean of the evaluated subjective category for training; Column 3 is the mean of the evaluated subjective category training (optimized) system; Column 4 shows the difference between test and training averages; and Column 5 is the statistical significance value obtained using the standard t-test.

Accuracy of the system improved in the test version: the results show that there is a statistical significance between the two versions ($p = 0.0360 < 0.05$). This result indicates that the optimized system can select better dialog strategies than the training system which randomly selects dialog strategies.

Table 5.2: Subjective Evaluation Categories: Training versus Testing

Evaluation Measure	Training	Testing	Δ	p-value
Accuracy	4.6435	5.3363	0.6928	0.0360
Likeability	5.7286	6.1830	0.4544	0.0928
Cognitive Demand	5.2000	5.3435	0.1435	0.6394
Annoyance	4.7596	4.1540	0.6056	0.0472
Habitability	4.000	3.8200	0.1800	0.6302

Likeability of the system improved slightly in the test version. As can be seen, both versions of the system have very high scores for likeability. It is possible to draw two conclusions: first the acceptance rate of the system is high; second, although the difference between the two versions is not statistically significant ($p = 0.0928 > 0.05$), the optimized behavior of the system provides more desirable interactions.

The mean values of *cognitive demand* and *habitability* are very close for the training and testing versions (see Table 5.2). Therefore p-values are not statistically significant. However, we can infer that the required cognitive demand is slightly higher than neutral level for both versions.

Habitability of the system is almost neutral (i.e. 4 in 7 point Likert scale) for both of the versions.

We believe that there is a connection between *accuracy* and *annoyance* categories, because if the number of re-asks and confirmation increases, the annoyance level might increase. For the test version, the reported annoyance level decreased and the result is statistically significant ($p = 0.0472 < 0.05$). Since the accuracy also increased for the test version, it might have had a significant impact on the decrease of annoyance.

5.3 Implications

Health screening and assessment dialogs are different than dialogs that are found in information-seeking applications usually studied by SDS researchers. The main goal of brief behavior change interview dialogs is to collect initial screening information, educate patients, increase their awareness about potential problem behaviors and, if needed, refer the patient to a treatment. This is usually the plan of standardized health interviews (e.g. [HHEA⁺10, SRC⁺05, NIA07a]) by national or international health institutions. So the system has to conduct the conversation according to that plan. The system usually needs to ask one question at a time and in a specific order, while the flow of the dialog adapts according to the received answers. The length of the dialog is also longer than current information-seeking dialogs.

Our work have several implications. Our reduced state space representation with multiple MDPs enables to learn approximately optimal dialog policies with a relatively low amount of data. Even though we designed the system for brief alcohol interventions, the approach that we use is easily applicable to any other similar health interviews (e.g. eating behaviors, exercising behaviors, use of drugs). Indeed, brief interventions are adaptable and useful for a variety of life-style related issues that target one specific problematic behavior.

Secondly, our collected dialog corpus will help the development of future data-driven research projects in the health domain.

Thirdly, we connect this work with the notion of intelligent virtual agents (IVA). Whereas we focussed our current discussion on our efficient approach for a spoken dialog-based interaction, our work is directly linked with our research on the graphical animation of the intelligent virtual agents that deliver the spoken intervention. In a recent study [LAYR13], we showed that empathic virtual agents that deliver computer-based behavior change interventions are much more engaging than the cur-

rently available text-only computer-based interventions. We created a model of empathic communication for an IVA to deliver behavior change interventions: in brief, the agent can sense the user’s facial expressions and answers, and adjusts its non-verbal responses accordingly (e.g. express concern or encouragement) to deliver its messages. Whereas there are debates about the impact of virtual characters communicating empathically with humans, our results showed that people are 31% more likely to use our empathic agent system compared to using the same intervention content delivered instead with text-only. We are currently in the process of integrating and evaluating our empathy agent model with the dialog manager discussed in this article.

Lastly, the performance of our system has also convinced medical and health-care personnel to conduct randomized clinical trials to evaluate health outcomes and potentially deploy our system in clinicians’ waiting rooms and community centers. Whereas computer scientists might think that the healthcare profession could be threatened by the creation of such virtual counselor technologies, they are instead quite enthusiastic about getting technological assistance to address some of the nations’ current epidemics (e.g. obesity, overweight, which put people at risk of a variety of chronic conditions such as diabetes, cardiovascular diseases, among others). Virtual counselors have many advantages, including increased accessibility to cost effective health interventions for people in need, increased anonymity and therefore self-disclosure of at-risk behaviors, which in turn leads to better healthcare, among many others [LAYR13].

5.4 Discussion

I created a spoken embodied conversational system which uses the Reinforcement Learning (RL) paradigm for dialog management. The system is able to learn dialog

strategies for initiative and confirmation selection. Our contributions to the SDS domain include the creation of a RL paradigm to the completely new domain of behavior change - where the dialog length is 4-5 times longer and where the nature of the dialog is less restricted than spoken dialog systems operated in tourist information domain.

We contributed to the health-care domain with the first system to use speech as an input medium with a RL-based approach. The initial evaluation showed that the dialog managers that are optimized with RL have the potential to reach optimal behavior, given enough training data.

The future research will involve extending our evaluation with more training data, and testing the optimized system with a larger number of subjects. Our system currently takes into account the best recognition of the speech recognizer. We plan to use partial observability concepts to deal with uncertainty, which stems from speech recognizer hypotheses: future versions may work with N-best speech recognitions instead of best speech recognition.

CHAPTER 6

Recognizing Named-Entities and Domain Specific Reasoning For Behavioral Health

Named-Entity Recognizers (NERs) are an important part of information extraction systems in annotation tasks. Although substantial progress has been made in recognizing domain-independent named entities (e.g. location, organization and person), there is a need to recognize named entities for domain-specific applications in order to extract relevant concepts. As described earlier my motivation is to develop a dialogue system to address some of the latest worldwide epidemics of behavioral issues (e.g. over eating, lack of exercise, alcohol and drug consumption). I focused on the domain of behavior change, especially *lifestyle change* because unhealthy behaviors are often associated with many health problems [Wil02]. For the dialogue system I developed a system to recognize of named entities related with the unhealthy behaviors has a crucial importance in language understanding. Recognizing named-entities gives possibility to integrate domain related knowledge to discourse. However, there is no named-entity recognizer designed for the lifestyle change domain to enable applications to recognize relevant concepts.

I describe the design of an ontology for behavioral health based on which I developed a NER augmented with lexical resources. My NER automatically tags words and phrases in sentences with relevant (lifestyle) domain-specific tags (e.g. [un/]healthy food, potentially-risky/healthy activity, drug, tobacco and alcoholic beverage). I discuss the evaluation that I conducted with manually collected test data. In addition, I discuss how my ontology enables systems to make further information acquisition for the recognized named entities by using semantic reasoners.

Because identifying behavioral concepts in text or user utterances requires some world knowledge, I created a behavioral health ontology to model world knowledge

for behavioral health problems, and a named-entity recognition system based on the ontology.

As discussed later, traditional named-entity recognition (NER) task focuses on finding and tagging proper nouns into predefined set of classes such as location, organization or person [TKSDM03]. In addition to these mentioned named-entities, recognizing numerical and temporal entities such as date, time, percentage, money have also been studied by researchers [FGM05].

By contrast, in order to address my focus on behavioral health, I am interested in extracting information about behavior-related concepts which are generally classified as common nouns. Specifically, my main goal is to classify named-entities into categories that are essential for the design and development of behavior health [Mat80] systems, which are mostly focused on lifestyle changes.

In the next section, I discuss latest research conducted in the named-entity recognition field and I compare domain-independent NERs against domain-dependent NERs. I then describe my general approach for the design and development of my behavioral health ontology. Finally I discuss the evaluation of my system and the current results that I obtained on a manually collected data set

6.1 Background in Named Entity Recognition

Ontology-based named entity recognition, annotation, and information extraction is used successfully in different domains including extracting relevant concepts in biological literature [MKS04] and the business intelligence domain [SFMB07]. In the food domain, Weigand et al. [WRL⁺12] designed a lexical resource for German, to perform relation extraction for recommending products and assisting online customers. A typical relation type is pairs of food items that are suitable to be consumed together. In addition to the relation types, each food in a relation is classified into healthy and

unhealthy categories. This system takes also into account context-dependent healthiness (i.e. having a medical condition such as allergy) which requires background information about a user.

Although ontology-based approach can be seen similar to using gazetteers (list of names of entities) approach in NER, the ontology approach provides additional advantages in terms of making further reasoning and knowledge acquisition for extracted concepts. I will discuss it in more detail in the following sections.

Traditional domain-independent named-entity recognition mainly concentrates on using supervised techniques to classify proper nouns into small number of predefined categories [TKSDM03], [NS07]. The disadvantage of this method is the need for collecting and gathering hundreds of labeled training data. Although there is available data for common categories (e.g. location, organization and person), for domain-specific categories it is not the case. Collecting and labeling hundreds of training data is not feasible for domains which deal especially with common names. Because common names (e.g., apple, gym, whiskey) do not have specific word-level features (i.e., orthographic information, orthographic patterns) as proper names (e.g., Apple, IBM, Henry Ford, 3M) which are used widely in supervised systems [NS07], the feature space for common names is a lot more restricted than the feature space of proper names.

Also document and corpus features including multiple occurrences, local syntax, and corpus frequency are not really useful for common names. Although these aspects are disadvantages in terms of using supervised techniques in recognizing common names, there is a possibility to use alternative approaches (e.g. lexical semantic networks, lexical ontologies) which are not directly available for proper name recognition, because there is no dictionary or lexical resource containing all proper names which are constantly being created.

Moreover, Krupka and Hausman [KH98] showed that using extensive gazetteers for proper name recognition does not really improve recognition accuracy.

The goal is to recognize health behavior related named-entities while user is interacting with the dialog system that I described before. Recognizing and classifying domain specific entities from utterances is the first step towards my the goal. Having an OWL ontology introduces additional possibilities for the recognized entities by using reasoners to classify them into further categories which has crucial importance in autonomous agent-based dialogue systems.

6.2 Approach

The NER uses 6 labels to tag recognized named entities. These are **(1)** healthy and unhealthy food labels for behaviors related with *diet*; **(2)** healthy and potentially-risky activity labels for *exercise* and *alcohol consumption* related behaviors (*activity* may involve alcohol such as partying); **(3)** healthy and potentially-risky place labels for exercise and alcohol consumption (*place* may have or involve alcohol such as night club); **(4)** alcoholic beverage label to recognize alcoholic beverages; **(5)** drug label to recognize *drugs*; **(6)** and tobacco label to label *tobacco products*. If the system can not find polarized label (e.g., healthy food, healthy activity, and potentially-risky place), it uses neutral labels (e.g., food, activity and place).

In addition to using ontology and modeling knowledge using the Web Ontology Language (OWL) [MvH04], the system is augmented with WordNet [Mil95a]. WordNet is used, if a concept does not exist in the ontology.

The advantage of my ontology-based augmented approach is twofold. First, it does not require to build and maintain extensive gazetteers because the system uses semantic network structure based on WordNet. WordNet-based named entity recognition has been implemented successfully for domain-independent NER by extracting trig-

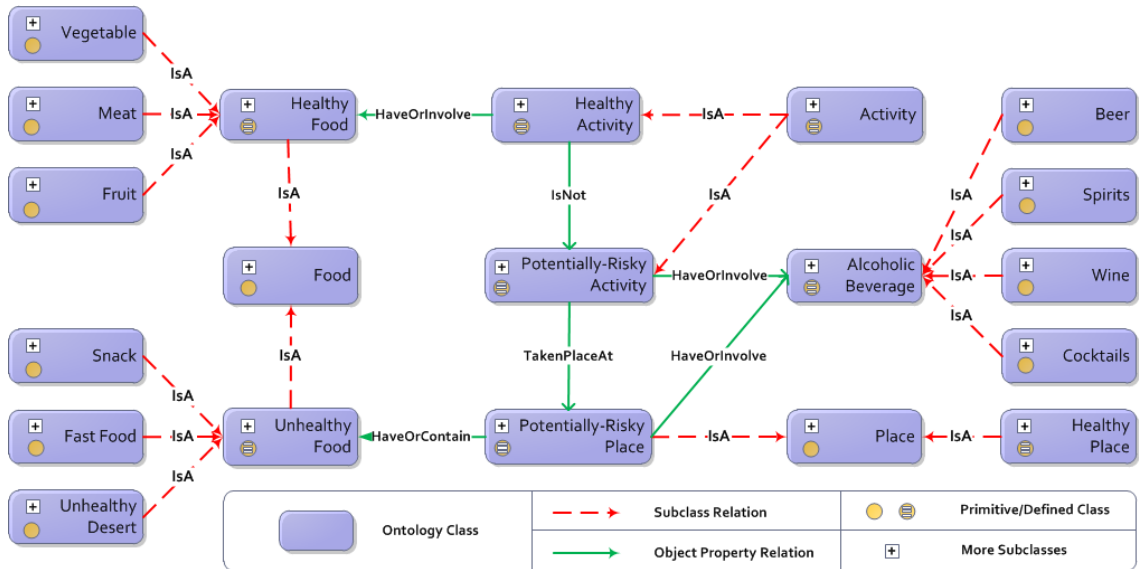


Figure 6.1: Behavioral Health Ontology

ger words from WordNet [MNPT02], and in video annotation applications based on semantic similarity [QGF10]. Second, my system’s domain is easily modifiable which makes it ontology-dependent but domain-independent. Therefore, I have adopted a different approach based on an extendable ontology model augmented with the WordNet. Although my NER can be only used for applications that are focused on lifestyle change, it is possible to use the application in different domains by changing the domain of the underlying ontology. In the next section I will discuss my ontology design.

6.3 Ontology Design

I have designed my behavioral health ontology in Protégé [KFNM04] with OWL 2. A small snapshot of my ontology is presented in Figure 6.1. There are 6 top level classes which correspond to categories of named entities. In total there are 60 classes, 490 individuals, and 20 object properties in my current behavioral health ontology.

Healthiness, unhealthiness and potentially-riskiness are defined to address most prevalent risky health behaviors such as obesity, excessive alcohol consumption, drug and tobacco use. For categorizing foods into the healthy and unhealthy categories, I considered following aspects based on [WRL⁺12]’s approach :

- Sugar content of the food item.
- Fat content of the food item.
- Artificial additives and extensive industrial processing of the food item.

For example *broccoli* is an instance of the *Vegetable* class which is the subclass of the *Healthy Food* class and *chocolate* is an instance of the *snack* class which is subclass of the *Unhealthy Food*.

Activity and *Place* concepts are defined in terms of *Healthy/Unhealthy Food*, *Alcoholic Beverage*, *Tobacco* and *Drug* concepts. I defined relationships between the ontology concepts by using the object properties such as *HaveOrInvolve* and *Taken-PlaceAt* (see Figure 6.1). As an example, a *Potentially-risky Activity* (e.g., drinking alcoholic beverage, eating junk food, smoking) may have or involve *Unhealthy Food* (e.g., fast food), *Alcoholic Beverage* (e.g., vodka) or *tobacco products*. I also defined the object properties which allows to perform knowledge acquisition between the *Healthy/Potentially-risky activity* and *Place* ontology classes. As an example *Potentially-Risky Activity* assumed to be taken placed at a *Potentially-Risky Place* (fast food restaurant). For a subset of the ontology structure which shows the relationships between ontology concepts (see Figure 6.1).

Individuals in the ontology structure represent instances of each class. For example, **grape** is an individual of **Fruit** class and transitivity between class structures implies that **grape** is also individual of **Healthy Food** class (fruit is a subclass of healthy food). In addition to entities with common names, for some classes I include proper names which are frequently used for some ontology classes. For example,

Burger King is an individual of **fast food restaurant** class which is a subclass of **Potentially-Risky Place** class because fast foods are generally classified as unhealthy. The **Alcoholic Beverage** class also contains many instances which has proper names (e.g. beer, vodka, whiskey brands).

I have also defined anonymous classes based on relationships between concepts using object properties. Object properties such as "haveOrInvolve" allow my system to make some inferences, including inferences which are not directly indicated based on the class hierarchy. For example "having or involving alcoholic beverage" is explicitly specified as *Potentially-Risky Activity*, if I query my ontology by using OWL description logic (DL) query with "drinking *some* Jack Daniels", it can infer that *Jack Daniels* is a *whiskey*, *whiskey* is a *Spirit*, *spirit* is an *Alcoholic Beverage*, and using *Alcoholic Beverage* is a *Potentially-Risky Activity*. Although it is not required for NER task, for applications that require additional information about the recognized entities, my ontology structure can be queried to retrieve taxonomic information about the entities. In the next section I will discuss WordNet which has a semantic network structure and used to augment my ontology.

6.4 WordNet

WordNet is a lexical database of English [Mil95a]. All word groups including nouns are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interconnected by conceptual-semantic and lexical relations. WordNet can be used as a lexical ontology based on hypernym/hyponym relationships between noun synsets. These relationship structure can be interpreted as super-class and subclass relationship as in ontology classes.

WordNet Distance: Semantic distance, similarity, and semantic relatedness are being used interchangeably by researchers and used in annotation, word sense

disambiguation, information extraction, information retrieval, etc. Since, there are different measures proposed for relatedness or distance [PPM04], it is important to distinguish these terms.

Budanitsky and Hirst [BH06] distinguish **semantic relatedness** as a more general concept of **similarity**. They attempt to demonstrate the difference between relatedness and similarity by an example: "Similar entities are semantically related by virtue of their similarity (bank-trust-company), but dissimilar entities may also be semantically related by lexical relationships such as meronymy (car-wheel) and antonymy (hot-cold), or just by any kind of functional relationship or frequent association (pencil-paper, penguin-Antarctica, rain-flood)." Therefore, similarity and relatedness does not refer to the same concept.

The semantic distance term generates even more confusion in terms of relatedness and similarity. Therefore, there are different approaches to calculate it. The semantic distance I am referring to is the distance in hypernym/hyponym tree. As I have mentioned before, WordNet can be interpreted as an ontology based on hypernym/hyponym relations. Thus distance between two words in hypernym/hyponym tree is more compatible with my goals for NER than relatedness or similarity concepts.

I used RiTa.WordNet¹ library to calculate semantic distance. The algorithm calculates the distance between any two senses of the two words (results is normalized within 0-1) with the specified Part-Of-Speech(POS) tag. For my purposes I use noun as POS tag. The algorithm (1) finds common parents of the two words, (2) calculates the minimum distance (shortest path) to the common parent from either of the words, (3) calculates the distance from the common parent to root of tree, and (4) normalizes the result (see Algorithm 2).

¹<http://www.rednoise.org/rita/wordnet>

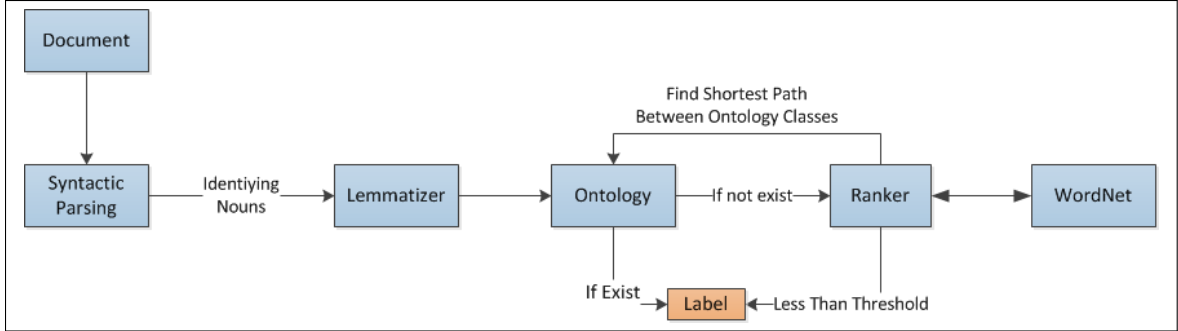


Figure 6.2: The System Architecture

6.5 Architecture

The NER performs tasks to locate nouns in the sentences based on the output of the Stanford Part-Of-Speech Tagger, then the identified nouns are lemmatized with the lemmatizer available in Stanford CoreNLP tool [TKMS03a]. The identified and lemmatized nouns are passed to the Tagger algorithm to be labeled into the following categories: (1) Healthy Food; (2) Unhealthy Food; (3) Healthy Activity; (4) Potentially-risky Activity; (5) Healthy Place; (6) Potentially-risky Place; (7) Drug; (8) Alcoholic Beverage; and (9) Tobacco. The system may use neutral labels from ontology for tagging, if the system can not identify polarized label for the named-entity (e.g. instead of unhealthy food, food label can be used).

Tagger algorithm (1) queries classes in the ontology, if it finds a matching class, it traverses the ontology to higher level classes to find a appropriate tag; (2) if the lemma is not equal to the name of any classes, it queries individuals in the ontology and finds the class of an individual (if the individual exists) and traverses the ontology to find an appropriate label; (3) a) if the noun does not exist in the ontology, it uses the distance algorithm (see Algorithm 2), the ranker component (see Figure 6.2) compares the distance between each class and the parameter noun, and then the tagger algorithm selects the class with the minimum distance to the noun; b) if the selected class is a first level class (e.g. Alcoholic Beverage, Drug/Narcotic) and the

distance is less than the threshold distance, it tags the name with the corresponding label; c) if the selected class is in lower position (e.g. Beer, Cannabis) in the hierarchy and the distance is less than the threshold distance, it tags with the corresponding tag.

The intuition behind using different distances for different level classes is as follows: if the minimum distance of a parameter noun is calculated for a lower level class in the ontology, it is expected that parameter noun is also a specific term, so the minimum distance to their common parent in hypernym/hyponym tree is expected to be short. For example, if the noun is *margarita* and the closest ontology class is *martini*, the expected distance is short because *martini* is a low level class in ontology. If *martini* did not exist in ontology and shortest distance to *margarita* is from *alcohol* class, the expected distance is longer than the distance of specific class because *alcohol* class is a high level class. For this specific example the distance between *margarita* and *martini* is 0.1 and their common parent is *cocktail* in tree hierarchy. The distance between *margarita* and *alcohol* is 0.3 where alcohol is the parent of margarita (common parent is alcohol too). Therefore using different threshold values for different level classes in the ontology helps to fine tune coverage of extension based on WordNet.

6.6 Data and Evaluation

Since there was no tagged data in behavioral health domain, I collected the test data manually from a variety of related websites which have relevant domains. For example, I have used meal recipe websites to find data related with the food domain. I have collected 88 sentences with 220 named-entities. Two annotators tagged the collected test data with the aforementioned labels. Then I performed two experiments with my NER. The recognition of an entity without healthy, unhealthy or potentially-risky label for the food, activity and place entities considered wrong. For example, if

Algorithm 1 Tagger

```
if Is parameter noun(pn) a class in ontology then
  Tag the noun with super class of corresponding class
else if Is pn individual in ontology then
  1.Find class Of the individual
  2.Find super class of individual's class
  Tag the pn with the super class
else
  1.Compare minimum distance between noun and Ontology classes by using Dis-
tance algorithm (See Algorithm 2)
  2.Select the class with shortest path to the noun
  3.If Selected class is first level and distance is less than higher-threshold
  return it as Tag
  4.If selected class is not first level and distance is less than lower-threshold
  return it as Tag
  Otherwise do not tag
end if
```

an entity recognized as *food* without having healthy/unhealthy label, I did not count it as a correct recognition. Since the *alcoholic beverage*, *drug* and *tobacco* ontology classes are all considered as unhealthy or potentially-risky in terms of behavioral health, the recognition of an entity in these categories does not need an additional label.

The experiment results are presented in Table 6.1. The first experiment was conducted using 0.2 as the distance threshold to the high level classes and 0.1 distance threshold the low level classes in the ontology. Second experiment was conducted using 0.3 as the distance threshold to the high level classes and 0.2 distance threshold to the low level classes. The performance of the NER is measured with $F_{\beta=1}$ rate:

$$F_{\beta} = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall}$$

with $\beta=1$ [Rij79]. Where, *Precision* is the percentage of the named entities which are correctly recognized by the system and *Recall* is the percentage of the named entities present in the test data that are recognized by the system. A recognized

Distance	Precision	Recall	$F_{\beta=1}$
0.1 and 0.2	83.32%	71.28%	76.80%
0.2 and 0.3	65.55%	81.44%	72.64%

Table 6.1: Behavioral Named Entity Recognizer Evaluation Results

named entity is correct only if it is the exact match of the corresponding entity in the manually tagged file. First, I conducted the experiment with the higher distance threshold values, so, the precision was low because of the false positives (unexpected results). The high number of false positive results were caused by the high threshold distances in the WordNet tree. I encountered many problems due to the word-sense ambiguity. For example *ice* and *glass* words were labeled as drug because *ice* and *glass* as a slang refer to a kind of drug. I observed many similar problems to this example in the first experiment.

Algorithm 2 Distance Algorithm

Locate the common parent of the two lemmas by checking each sense of each lemma

if No common parent found **then**

return 1

else

1. Calculate min distance to common parent (the shortest path from either lemma to common parent)

2. Calculate distance from common parent to root (length of the path from common parent to the root of WordNet ontology)

3. Calculate and return the normalized distance to common parent as:
 ($minDistToCommonParent / (distFromCommonParentToRoot + minDistToCommonParent)$)

end if

In the second experiment, the precision increased significantly while recall decreased. It was the result of the low threshold values for the distances. In this experiment the number of false positives decreased significantly while false negatives (missing results) increased. It was the result of the decreased coverage of the system

due to the low distance threshold. I did not observe as many unexpected results as in the first experiment because of the slangs but I observed an increase in unlabeled named-entities.

Another factor which affects the results is the output of the part-of-speech tagger and lemmatizer. Although they worked with high accuracy in general, for some cases they did not give expected output.

Overall, although precision, recall and F-Measure results are not high, the results are acceptable for the first version and promising for the future versions.

6.7 Discussion

I designed a named-entity (NE) recognizer for the lifestyle change domain. I addressed the differences between traditional NE recognition and the domain specific NE recognition. To address my problem in recognizing lifestyle related entities in text, I designed a behavioral health ontology. Based on my ontology model, I created a named entity recognizer. Also, I identified other possible use-cases of my ontology. To extend the ontology for the named-entity recognition purposes, I augmented it with the WordNet. I used a hypernym/hyponym tree and calculated distances between synsets.

I conducted two experiments with different distance threshold values and reported the results. I observed that threshold distance has a significant effect on precision and recall. While high threshold values increase the recognition rate, it causes unexpected false positives because of wrong labels. I believe that I can address this issue by using dynamic distance threshold for different ontology classes in future. Although resulting precision, recall and F-Measure results are not high, they are acceptable for the first version and promising for the future versions. In the future versions of my named-entity recognizer, I will conduct experiments with the dynamic distance thresholds.

Finally, the NER is available to use in most prevalent behavior related health applications. It can enable to recognize domain-related named entities in user utterances or any kind of text. In addition, making inferences is very important for dialog systems, it enables to perform inferences based on recognized concepts and relationships between concepts.

CHAPTER 7

Lexical Resources & Semantic Networks

Discussing issues about at-risk behaviors such as heavy drinking are highly emotional for people to talk about (e.g. shame, discouragement, anger, hopefulness, satisfaction, pride). Empathy and positive regard toward the client are therefore critical therapeutic conditions to create an atmosphere of safety and acceptance where clients feel free to explore and change [MR09b]. The sentiment and emotion recognition is first step of understanding conveyed emotion and act appropriately. To recognize conveyed emotions and sentiments from user utterances I create lexical resources for prior sentiment polarities and affective ratings of words. I also created general purpose unified semantic network (USN) which is augmented with affective information and ontologies. The sentiment and affective ratings lexicons are created by exploiting expert knowledge stored in USN. I discuss details starting from next section.

7.1 Lexical Resources & Semantic Networks

Identifying innate or prior sentiments and affective ratings (i.e. valence, arousal, and dominance) of words is a challenging task. It has been studied by a variety of research groups in the last decade. In this section my approach relies on expert knowledge to compute prior sentiment polarities and affective ratings of words. I unified existing lexical resources by using a graph-database. The connected nature of lexical resources makes it possible to use linked data representations. Recent advancements in graph database systems have demonstrated that it is possible to develop highly scalable graphs. There is a substantial amount of work conducted in the development of semantic networks and specific purpose lexical resources (e.g. sentiment or affect lexicons) which contain explicit and accurate information. I am interested in leveraging existing resources to derive new knowledge by exploiting the potential information

based on connections between lexical resources. In conjunction with capabilities of graph databases, relationships between distinct lexical resources provide a considerable potential for exploiting existing relational information by using graph queries and graph-based inference algorithms. I built a new graph-based semantic network from existing resources. I designed new inference algorithms that can exploit the potential knowledge that exists between interconnected semantic networks. My particular interest is sentiment and affect recognition. Unified Semantic Network (USN) contains affective ontological information. I designed new graph traversal algorithms to infer sentiments and affects that words convey. I evaluated the results against well-known lexical resources for sentiments and affects. The results show that it is possible accurately infer prior sentiments and affective ratings conveyed by words. Based on my approach, I provide new highly-accurate lexicons for sentiments and affect ratings for 75,658 words, made available to the research community. In addition, USN can be used for general purposes, such as inference about word similarity and relatedness, by using some of the proposed algorithms or by creating new algorithms. I also provide full access to USN for research purposes. USN is likely to have many implications, not only in terms of development of new domain-specific lexicons by inference, but also in discovering new information for words using the rich linked data representation that brings together many lexicons and semantic networks.

7.2 Introduction

The development of semantic networks and lexicons is a very laborious process, usually takes years of intensive human effort. In the last half-century a substantial number of resources developed which relies on expert knowledge [SDS66, Mil95b, Osg57, BL99]. However, not much work has been conducted to link available semantic networks and lexicons to derive highly accurate knowledge by exploiting potential

semantic information in the connections between distinct resources to address a variety of needs, which is my current approach. Graph-databases are well positioned to provide the required infrastructure, not only for link-data representation, but also for information exploitation from multi-source data.

A high number of lexical databases, specific purpose word lists, and semantic networks have been developed in the last decades. They can be linked by using graph-databases which introduces new possibilities, not only for performing new inferences, but also for discovering new insights that are induced by multiple sources. Some of the current resources are just designed to address specific important problems such as lexicons for sentiment polarities [WWH05, ES06, GGT13, HL04], or lexical resource for affective knowledge [SV04, NPI09, BL99, WKB13], which are also my focus in this paper. The main well-known limitation of the domain-dependent, manually-created lexical resources is that, for words that do not exist in the lexicon, it is not possible make any inference or estimation about relevant aspects [SV04, BL99, WWH05]. However, these contain accurate information that can be used as expert knowledge or gold standard. The comprehensive domain dependent lexicons which are created automatically usually contain noisy information [ES06, GGT13]. In construction of the graph, I use manually-created lexicons for the sake of accuracy.

Some of the lexical resources such as FrameNet [BFL98], WordNet [Mil95b], General Inquirer [SDS66], ProbBank [PGK05], and VerbNet [Sch05], target more general and less constrained problems than specific purpose lexicons. Each general purpose lexical resource aims at providing different insights about lexical entities. For example, WordNet is indisputably the most popular lexical resource that concentrates on providing ontological information such as hyponymy and hypernymy for words that are grouped as synonym sets. While these aspects are very important for natural language processing applications, they don't provide domain-dependent insights. Since

WordNet is a semantic network, it allows to perform inferences that are not directly available in WordNet such as word similarity and relatedness [HSO98]. However, the main shortcoming of WordNet is the sparsity of the connections between synonym sets. That is also valid for other general purpose semantic networks [Sch05, BFL98]. I use WordNet [Mil95b] and General Inquirer [SDS66] as a backbone of my graph-database, which enable to exploit multi-source information.

I have three direct contributions to the field. The first contribution is the creation of unified expandable semantic network which connects multiple semantic networks and lexical resources as a core graph and make possible multi-source information exploitation. My unified expandable semantic network is easily expandable and adaptable to new domains, and makes it possible to populate new lexicons from small lists. I discuss it in Section 7.5. The second contribution is the creation of new sentiment lexicon for prior polarities of words for all alphabetic entries in WordNet which consist of 75,568 words (i.e. I excluded alphanumeric entries and phrases). The tests show that my results are better than SentiWords [GGT13] which can be regarded as state-of-the-art and most comprehensive lexicon for prior polarities of words that is generated by using SentiWordNet [ES06]. Third I generated a new lexicon for affective ratings of 75,568 words, which is one of the most comprehensive lexicon generated so far for prior affective ratings of words. The currently available affective rating lexicons are created manually by experts [BL99] or by crowd-sourcing methods [WKB13]. The results show that my affective ratings have high correlation with the expert knowledge or crowd-sourced lexicons.

In the remaining of this section, I discuss related work, the structure of the graph, my approach for generating sentiment and affective rating lexicons, and the evaluation of the results.

7.3 Background on Lexical Resources

The idea of unifying lexical resources is not new, it has been studied before by some research groups. For example, Lei et al. [SM05] combined WordNet [Mil95b], VerbNet [Sch05], and FrameNet [BFL98] to create richer knowledge-base for robust semantic parsing. In another unified network called BabelNet, lexicographic and encyclopedic knowledge from WordNet and Wikipedia were integrated [NP10]. UBY is defined as a large-scale unified lexical-semantic resource by its creators [GEKH⁺12], that gathers together 9 lexical resources in English and German and includes WordNet, Wikipedia, Wikitionary, FrameNet, and VerbNet as English resources. The goal is usually enriching available data by using other resources, and it is not possible to exploit individual connections between words and concepts. In my system, I provide the opportunity to make multi-source inferences, rather than simply focus on enrichment for labeling the words. In addition, I provide an environment which enables the use of an SQL-like expressive graph query language called Cypher, and graph traversal algorithms (e.g. shortest path). I use WordNet as a core of USN which is usually the core of all unified resources. Therefore WordNet can be used an interface to integrate unified resources to my semantic network such as [SM05, NP10, GEKH⁺12].

One of my main focus is exploiting prior sentiments of words. There are many sentiment lexicons that have been built in the last decade [WWH05, ES06, GGT13, HL04]. Some of them are manually created, and therefore they are not comprehensive but accurate (e.g. MPQA [WWH05] and Bing Liu’s opinion lexicon [HL04]). Some of them are automatically created, so that they are comprehensive, but they contain noisy information [ES06, GGT13]. For example, SentiWords [GGT13] is automatically generated based on SentiWordNet [ES06] for prior polarities of words. SentiWordNet [ES06] is generated by automatically annotating all WordNet synsets according to their degrees of positivity, negativity, and neutrality, by using semi-

supervised methods. General Inquirer (GI) [SDS66] is another lexical resource which contains word lists for sentiments in addition to other purposes: it is accurate but not comprehensive. Unified Semantic Network (USN) relies on MPQA [WWH05] and General Inquirer [SDS66] for sentiment inferences. Sentiment lexicons are freely available for research purposes but there are exceptions [PFB01]. My goal is creating a free, comprehensive, and highly accurate lexicons that are generated automatically based on other manually created lexicons.

I also aim at creating a comprehensive lexicon for affective ratings. Affective ratings have three components: valence (the pleasantness of the stimulus), arousal (the intensity of emotion provoked by the stimulus), and dominance (the degree of control exerted by the stimulus). So far, nearly all research has been conducted based on Affective Norms of English Words (ANEW) [BL99] for 1,034 words. Warriner et al. [WKB13] has extended the database to 13,915 words by using crowd-sourcing. Bestgen et. al. [BV12] has also used ANEW and extended the database to 17,350 words by using a corpus-based approach. While [BL99] and [WKB13] can be accepted as gold standards, Warriner et al. [WKB13] has low correlations with human rated affects. General Inquirer (GI), which is integrated with USN, contains affective categories for valence, arousal and dominance, based on Osgood's three semantic dimensions [Osg57]. I use GI to populate affective ratings of words and, I use the gold standards mentioned above for testing the results. In USN, I also include WordNetAffect [SV04] which is another resource that mostly contains ontological affective information (rather than an extensive list of words).

The above lexicons are useful for a wide range of tasks, but they are fixed resources. Building lexicons can have three benefits, **1)** much larger lexicons can be developed inferentially; **2)** different dimensions of sentiment that might be pressing for specific tasks can be captures; **3)** it is possible to develop lexicons that are sensitive to the

norms of specific domains. I can develop lexicons that are sensitive to the norms of specific domains. Three most popular approaches are possible to create new lexicons, simple WordNet propagation, weighted WordNet propagation, and semi-supervised methods. In simple WordNet propagation, the algorithm starts with small number of seed words, and then populates the list by using WordNet relations [AB06, GSS07]. In weighted WordNet propagation, the algorithm uses propagation, not only for the senses of the original seed set, but also attaches scores to words, reflecting their intensity [BGHM⁺08]. Supervised and semi-supervised approaches are also widely used for lexicon creation [ES06, BV12]. My approach relies on exploitation of multi-source lexical resources by using graph traversal algorithms, and the details of my approach is described in Section 7.7.

7.4 Storing Semantic Data As A Graph

Graph databases offer several key advantages against alternative methods, such as using relational databases, or semi structured text files and spreadsheets. They provide scheme-less and efficient storage for semi structured data. Semantic networks can be seen as an arbitrarily connected set of points, so that they are semi-structured. The relationships between concepts are beyond the capability of referential integrity concept which is used in relational databases usually. The conceptual relationships contain many irregularities and require that they be stored with rich relationships among concepts. This makes it impossible to fit them in a scheme of a tabular structure. The possibility to express semantic concepts and their relations in a similar way to how we visualize them when communicating the ideas, removes obstacles in performing deep inferences easily, and also allows for their storage to be similar to the way we think about them. Thus graph databases are regarded as white-board friendly.

Graph databases create new opportunities for the querying the semantic networks which was not possible, or very hard to express, in prior storage methods. It is possible to express queries as traversals in addition to SQL-like queries. Graph databases are fast when it comes to query linked information: it is possible to perform fast deep traversals instead of slow SQL queries that span many table joins. It is very natural to express semantic networks related problems (e.g. distance between two synonym sets) with traversals.

While we represent a semantic network visually, we naturally draw some nodes to represent the words and edges to represent relations, which is basically a simple graph database. The graph database paradigm provides increased expressiveness, scalability, functionality to perform complex queries, and also enables to make inferences using graph algorithms (e.g. Dijkstra, Shortest Path). The graph databases have recently become very popular in the social networking area as they facilitate social network analysis. The algorithms that are used in social networks are also usable in semantic network analysis to perform inferences about words.

To increase the expressiveness of the semantic networks and perform multi-source inferences, I created USN. The unified knowledge enables to ask interesting questions, and to go beyond simple inferences by exploiting inter-network relations.

Graph databases are usually compared to RDF triples. While it is possible to store RDF triples in graph databases and run SPARQL queries on them, it might not be possible every time to transform graph databases to RDF triples. There are some subtle, but important, differences which stem from graph morphisms. Graph databases can store various types of graphs, including undirected or directed graphs, weighted graphs, hyper-graphs whereas RDF triple can only store edge-labeled graphs. Edge-labeled graphs provide the ability to denote the way in which two vertices are related that is not sufficient to represent all relationships in a multi-source semantic

network. The ability to naturally represent different graph morphisms enable to unify multiple resources. For example, WordNet relations are directed, to represent General Inquirer categories, hyper-graphs are needed which allow an edge to connect to an arbitrary number of vertices. To represent affective ratings, I use weighted graph. Therefore, my semantic network contains multiple types of graph together.

I created the graph database using Neo4j¹ platform. Neo4j has a query language called Cypher². The database can be queried with a web-based interface and the results can be visualized (an example showed in Figure 7.2). The language drivers are available for most of the mainstream programming languages including, but not limited to, Java, .NET Framework languages and Python.

Neo4j can scale to a few billions of nodes, relations and properties in a single machine [HP13]. It is possible to expand the current database to multiples of its current size, therefore there is room for expanding it for specific needs.

7.5 Core Structure of Unified Semantic Network

There has been a lot of work done in the creation of semantic networks and lexical resources in the last decades. These resources help to develop many applications, usually by requiring extra effort to adapt, or reuse, to build new lexical resources according to the domain of the application. In this section I discuss my approach in building the unified lexical resource from existing resources. The resource will lead to development of new approaches based on semantic relations. USN has the flexibility of being extended for new lexical requirements, as well as being queried with SQL-like query language.

¹<http://www.neo4j.org/>

²<http://www.neo4j.org/learn/cypher>

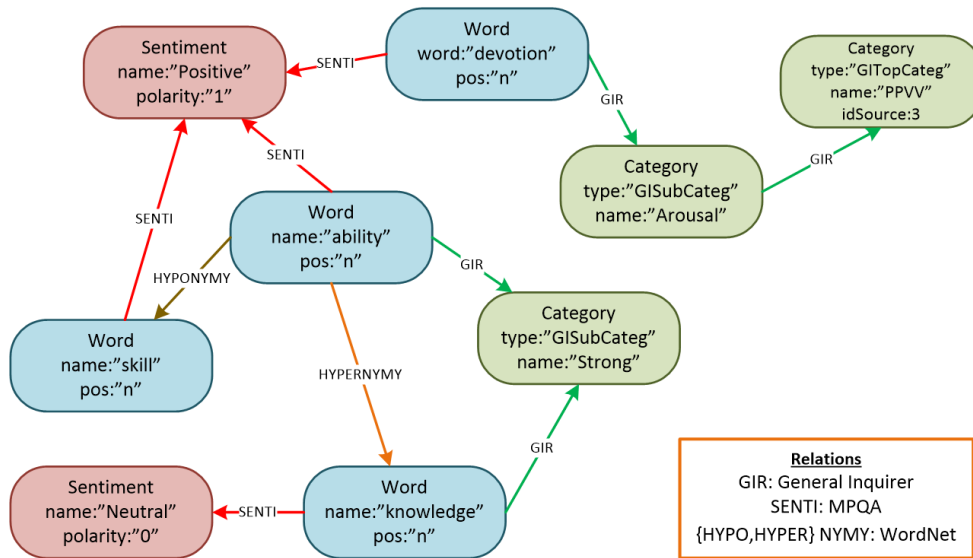


Figure 7.1: Sample of Structure of the Graph

I started to build USN by using WordNet 3.1 [Mil95b]. I grouped the different senses of each word by using word and part-of-speech tag pairs. I removed all entries in WordNet which do not convey any semantic meaning such as numeric entries (100st, 11), alphanumeric strings (e.g. 1-dodecanol, BLU-82) and phrases (e.g. phase space, physical process). I included 91,984 distinct word and part-of-speech tag pairs. I included all the WordNet ontological relationships (i.e. *HYPERNYMY*, *HYPONYMY*, *MERONYMY*, and *HOLONYMY*) with their original names as graph edges. I augmented it by using General Inquirer [SDS66]. General Inquirer (GI) contains 182 categories which are grouped under 26 groups. There are 81,988 relations between words and General Inquirer categories, which are denoted by GIR (General Inquirer Relation) type relations. For example, *EMOT* category contains emotion related words, *Pleasur* category contains words that are indicating the enjoyment of a feeling, including words indicating confidence, interest and commitment. Both

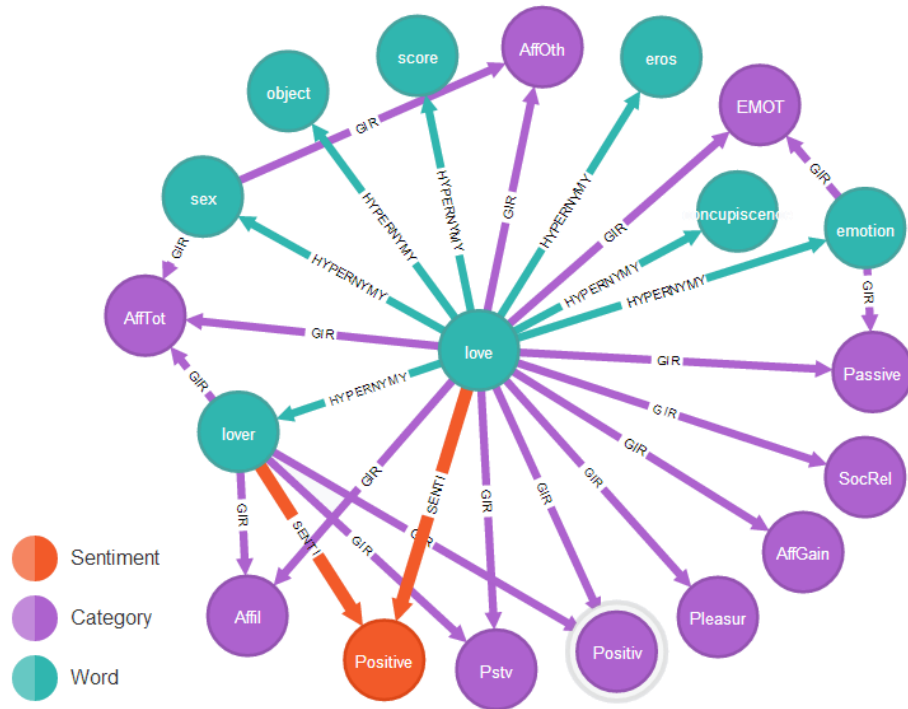


Figure 7.2: A snapshot from query result visualizer which shows relations of word *Love* in WordNet, General Inquirer and MPQA. Arrow labels shows the relationship type. GIR is General Inquirer Relationship. SENTI is sentiment relationship. HYPERNYMY is WordNet hypernym relationship.

EMOT and *Pleasur* are grouped under the parent category "*pleasure, pain, virtue*" as depicted in the Figure 7.1. The coverage of GI is not limited to affective knowledge. It contains comprehensive categories, including but not limited to Osgood tree semantic dimensions [Osg57] (e.g. strong-weak, positive-negative, active-passive), motivation related word categories (e.g. Need, Goal, Try, Persist), cognitive orientation word categories (e.g. Think, Know, Perceive), among others..

Each word can fall under multiple categories which describe a word in terms of multiple aspects. In total, GI contains 11,790 words. I grouped each word and part-of-speech tag pair, as I did for WordNet. In the graph, I included approximately 10,191 word from GI. I created a category node in my graph for each GI category,

and linked each word (which was previously acquired from WordNet) in that category to the category node.

7.6 Affective Knowledge

The prevalent affective information in text-based communication started to gather a lot of attention from natural language processing researchers. There is wide interest to develop future human computer interaction applications which take into consideration a user's emotions [KPJD13], motivation, frustration [MKC10], attention and engagement [CLH12, CHH12, PHA⁺10]. As it has been discussed, there has been a substantial effort to create lexicons for affective knowledge [SV04, NPI09, WWH05, ES06, HL04, GGT13, BL99, WKB13].

The domain of an application has an effect in determining what type of affective information is needed. For example, sentiment polarity information is relevant for movie reviews [SPW⁺13a, PL08], whereas for e-learning systems, attention and frustration information is more relevant [CGSG04a]. However, the lexicons created up-to-date only take into account Ekman's 6 basic emotions (i.e. happiness, sadness, fear, anger, disgust and surprise) [Ekm92] that are usually very limited [SV04], or affective ratings (i.e. valence, arousal or dominance) [BL99, WKB13] which are more comprehensive but not enough, or sentiment polarity lexicons [WWH05].

The limitedness of the emotional labels used in categorical approaches (e.g. Ekman's 6 basic emotions) usually fails to describe the complex range of emotions that can occur in text-based communication. To address the limitations of distinct emotion labels, the dimensional approach can be used to represent emotions [Rus79, Whi89]. Because this approach represents emotions using 2 or 3 dimensions (valence and arousal, and sometimes the third dimension of dominance), it is also more compatible with the dimensional meaning representation of words [Osg57]. The more recent

emotion models also use multi-dimensional approaches, such as the Plutnick's wheel of emotions which consists of 8 basic emotions and 8 advanced emotions each composed of 2 basic ones [Plu01].

Since the graph structure allows to represent each word in terms of multiple aspect, it is possible to express the emotions that a word convey using multiple dimensions. Therefore, I decided to incorporate Parrott's tree structured list of emotions [Par01], which is comprehensive hierarchical list of emotions. It contains multi-dimensional emotional models as well as the basic emotion categories. To bridge the gap between Parrot's emotional tree structure and other models, I used affective ratings that are available in general inquirer [Osg57]. I also integrated WordNet Affect [SV04] as a lexical resource, which contains list of words for Ekman's six basic emotions. The new lexicons can be easily generated by the proposed algorithm in Section 7.7 for fine-grained emotions of Parrot's emotion model.

For sentiment polarities of words and subjectivity scores, I integrated MPQA lexical resource [WWH05] which contains approximately 8,222 words. For sentiment polarities of words, GI also provides a list.

Emoticons used for symbolic representation of a facial expression have become standard way to convey feelings in text-based communication [Lo08, WD01]. Emoticons are used frequently in recognition of sentiments and emotions from texts [GBH09, KWM11, Rea05]. I linked emoticons from multiple sources that are popular, such as Yahoo messenger, Skype and Facebook. I associated each emoticon with its polarity and related emotions. The users of USN can query and find out semantic valence and conveyed emotion using standard emoticons.

I also linked emotions that are specifically experienced during a learning experience. According to several studies ([BDRG10, CGSG04b, GDC⁺08, WK06]), there are negative and positive prominent emotions experienced during problem solving,

reasoning and comprehension of technical material. I incorporated engagement and disengagement categories to the graph (see Figure 7.1), which is useful for applications in tutoring and e-learning domains [GD11, CGSG04a]. Disengagement is usually related to the negative affect states of frustration, boredom, anxiety and confusion, while engagement is usually linked to the positive states of delight, curiosity and surprise. Most of these emotions are included in Parrott’s ontological tree structured list of emotions [Par01]. I simply defined engagement and disengagement in terms of other emotions. In addition, I used the categories (i.e. valence, cognitive orientation, motivation first level categories) from GI to define learning related emotions. Adding new concepts to graph database is most of the time as easy as reusing existing knowledge to define new one.

7.7 Creating New Lexicons for Prior Sentiment Polarities & Affective Ratings

As I described in previous sections, USN contains richly connected multi-source accurate data. My hypothesis is that by exploiting the connections between the multi-source data, I can generate highly accurate and comprehensive lists for variety of purposes. My special focus area is sentiment and affective information. I designed an algorithm called ”discounted shortest path” to exploit the information stored in the graph.

As a background, I explained before the resources that I use in the graph. One of them is WordNet that contains ontological relationships between the synsets. Other one is General Inquirer which contains large number of categories including but not limited to sentiment polarities and affective rating word lists. I also integrated MPQA to the graph, it contains list of words for prior sentiment polarities. In the lowest

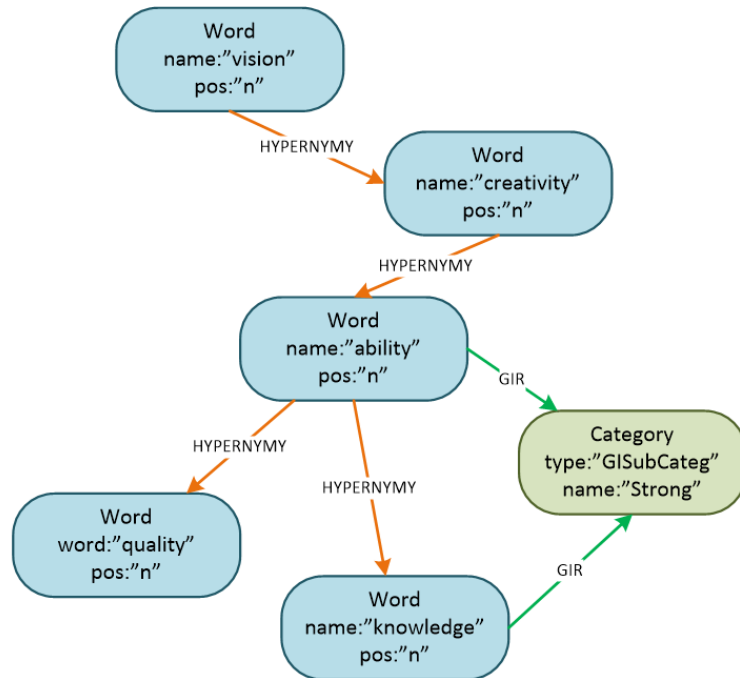


Figure 7.3: An example from the graph

layer of the graph, there are word and part of speech tag pairs (i.e. word nodes) and there are relationship between them, this structure is basically WordNet. On top of it, there are connections to GI and MPQA categories. For example, I have 3 categories (or nodes to represent categories), positive, negative, and neutral for MPQA. Each category is connected to the words in WordNet level. For example, in MPQA there are several thousands of words in positive category, for each of them I establish a link between MPQA positive node and word-part-of-speech tag pair. In my algorithm I exploit information from these 3 resources.

My algorithm briefly, receives a word and pos tag pair and calculates score of positivity, negativity, and neutrality for prior sentiment polarities and calculates score for valence, arousal, and dominance for affective ratings.

Discounted Shortest Path algorithm receives a word and returns a score which represents discounted score for destination category. For example, if a word is in MPQA positive category (e.g. vacation), it has a direct connection to positive MPQA

node, the distance is one, the score is also one, there is no need to discount. However, if there was no connection from queried word to destination, the system tries to find shortest path to destination by traversing hypernym and hyponym links of WordNet, and also to find categorical information, it uses Senti link to connect MPQA. The discounting is performed only for WordNet links. For example, if the shortest path is 4 which contains 3 WordNet links and 1 MPQA link, degree of discount factor becomes 3. The intuition is, there are around 8000 words in MPQA which has score 1 for positivity, negativity, or neutrality. Hypernym and hyponym of these words are not directly connected to MPQA sentiment categories, but connected via WordNet links, for each WordNet link, I discount because the distance increases.

As an example in Figure 7.3, I demonstrate simple discounting for dominance score calculation. GI Strong category is on of the Osgood’s three semantic dimensions [Osg57] which contains words relating with dominance affective rating. The word vision is not directly available in the category but multi-source structure of the lexicon makes it possible to compute discounted score. In formula 7.1, γ is discount coefficient and n is the number of WordNet links on shortest path to a target graph. For the case of **vision** word, the target category is **Strong** GI category and number of WordNet links on shortest path are two. For the affective ratings, I use 0.95 as discount coefficient, for sentiment scores I used 0.90 as discount coefficient.

$$p = \gamma^n \tag{7.1}$$

I execute discounted shortest path algorithm 4 times for positivity to find path to MPQA Positive and GI Positive categories by traversing hypernym and hyponym relationships and I calculate the average. For negativity score, I do same but the destination is MPQA Negative and GI Negative nodes. For neutrality score, I execute

it two times for MPQA Neutral category by traversing hypernym and hyponym links. The reason is there is not a neutrality category in GI.

It is sometimes not possible to calculate the shortest path, because there might not be a path from the word that I try to find sentiment for. In those cases, I calculate top 10 conceptually similar words to the input word by using GI categories. The conceptual similarity is calculated by finding the number of common GI categories, two or more words belong to, and limiting to top 10 words which have highest number of shared GI categories. For example, I calculated the top 10 conceptually most similar words to “*frustration*”, I get following result “*frustrated, anguish, guilt, hurt, anxiety, distress, upset, disturbed, disappointment, scared*”. I perform this 2 times, in first time I find set of 10 similar words to the input, in second turn, I find the set of 10 most similar word to each of the 10 words that I find in first turn. In addition, I use synonyms from WordNet. Then I try to calculate discounted shortest path to target category for each word and calculate the average. That gives us the score for positivity, negativity, or neutrality and affective rating scores.

I compute prior polarity labels from positivity, negativity, and neutrality scores. I use two formulas that are written below f_m and f_d . The formula f_m computes absolute value of twos scores, f_d computes difference between them. Each formula yields accurate result against different lexicons (see Section 7.8). These formulas are used in SentiWords to calculate prior polarities from posterior polarities of SentiWordNet,

$$f_m = \begin{cases} f(PositiveScore), & f(PositiveScore) > f(NegativeScore) \\ -f(NegativeScore), & otherwise. \end{cases}$$

$$f_d = f(PositiveScore) - f(NegativeScore)$$

7.8 Evaluation

I perform 2 evaluations to test accuracy of the designed algorithms. First to test the generated sentiment lexicon against other automatically generated comprehensive prior sentiment polarity lexicons. Second I performed tests for testing affective ratings.

7.8.1 Prior Sentiment Polarities

I generated new sentiment lexicon based on the discounted shortest path algorithm (see section 7.7). I evaluated prior polarities by comparing SentiWords which is recently published lexicon derived from SentiWordNet. SentiWordNet contains posterior polarities of words for each senses of the words, that is semi-automatically created. SentiWords is very comprehensive, it contains prior sentiment polarity for each word-part of speech tag pair. I conducted two tests using the f_m and f_d formulas that are mentioned above. I get %63.8 correlation with SentiWords, when I use f_m . When I used f_d , I get %85.5 correlation with SentiWords. While SentiWords is very comprehensive it is not the gold standard, it is automatically generated from another semi-automatically created lexicon. I used Bing Liu's opinion lexicon (4685 negative and 1978, total 6663 words) which is annotated by humans, that can be used as a gold standard. I conducted two tests based f_m and f_d formulas. I get %72.4 for negatives and %75.7 for positives accuracy with f_d formula. I get %86.6 for negatives and %86.0 for positives accuracy with f_m formula. I also tested SentiWords against Bing Liu's opinion lexicon. Prior polarities in SentiWords matched %40.1 of the negatives words, and %42.1 of the positive words.

I generated 2 lexicons with f_m and f_d which are available online³. Since f_m performed better against gold standard, it contains more accurate prior polarities. My sentiment lexicon contains 75658 which is basically all alphabetic word entries in WordNet (lemma and pos combinations).

7.8.2 Affective Ratings

I evaluated the discounted shortest path based on Norms of valence, arousal, and dominance for 13,915 English lemmas [WKB13]. In this lexicon, there are scores between 1-10 for each affective rating category. The algorithm calculates scores but not in same scale. I make discreet 2 categories for each affective rating, high and low. I divided highest scores to 2, and scores below half of highest score are labeled as low and scores above half of highest score are labeled as high. For the calculated scores, 0-0.5 is labeled as low, 0.5-1.0 is labeled as high. The test yields for valence %75.7, for arousal %82.2, and for dominance %72.2 of the time same labels with the gold standard. I generated a new lexicon which contains 75658 word part of speech tag pairs. It is most comprehensive affect ratings lexicon that is created up-to-date. I used categories (i.e. Osgood's three semantic dimensions [Osg57]) available in General Inquirer to generate this lexicon. The results are acceptable but I believe that if same method will be used to generate lexicon in conjunction of gold-standard, it might yield more accurate result. I did not do that because I want to test the system against the gold standard. Finally, the affective ratings lexicon is available to use for research purposes.

³usn.cis.fiu.edu

7.9 Discussion and Implications of Unified Semantic Network

I created a graph-based unified semantic network from multiple general purpose semantic networks and lexicons. I augmented them with affective lexicons and semantic networks. I demonstrated the potential of multi-source information exploitation by testing the results against well-known lexicons. I created new comprehensive lexicons for prior sentiment polarities and affective ratings. Affective ratings lexicon is the most comprehensive one that is created to the best of my knowledge. Prior polarities lexicon is also comprehensive and the results show that, it yields more reliable prior sentiment polarities than currently available state-of-the-art lexicons. The both lexicons that I created are available for research purposes. In addition, the graph database will be fully available to use for research purposes. It can be accessed in the following web page⁴.

⁴usn.cis.fiu.edu

CHAPTER 8

Sentiment Analysis

Associating words with prior sentiment labels and affective ratings does not help to fully understand conveyed affective information. To address that I developed a sub-module that can enable to recognize posterior sentiments in short sentences such as user utterances or news headlines. The sentiment recognizer takes into account interactions between words by using compositional semantics. Together with behavioral health named-entity recognizer (described in Chapter 6), lexical resources (described in Chapter 7), and sentiment analysis module can provide complete solution for emotion polarity recognition.

There has recently been growing interest in sentiment polarity (or valence) and emotion sensing using a variety of signals. Text, as a communication channel, gathers a substantial amount of interest for recognizing its underlying sentiment (valence or polarity), affect or emotion (e.g. happy, sadness). I consider recognizing the valence of a sentence as a prior task to emotion sensing. In this section, I discuss my approach to classify sentences/user utterances in terms of emotional valence. The supervised system performs syntactic and semantic analysis for feature extraction. My system processes the interactions between words in sentences using dependency parse trees, and it can identify the current polarity of named-entities based on on-the-fly topic modeling. I compared the performance of three rule-based approaches and two supervised approaches (i.e. Naive Bayes and Maximum Entropy). I trained and tested the system using the *SemEval-2007* affective text dataset, which contains

news headlines extracted from news websites. The results show that the systems that I designed outperform the systems demonstrated in *SemEval-2007*.

8.1 Introduction

In the last decade, emotion and sentiment analysis research has become a highly active field due to the increased necessity to recognize emotions, sentiments, opinions or affects conveyed through text. The possible applications which might benefit from recognizing affective information acquired from text, include but are not limited to sentiment analysis for customer reviews [PL08], opinion mining [LZS09], reputation management systems [YN05], affective and natural language user interfaces, such as spoken dialogue systems [THSh⁺11, YALR13]. My system concentrates on sentence level emotion polarity recognition (i.e. positive or negative), which I consider as a prior task to emotion recognition.

I use SemEval-2007 Task 14 affective text dataset, to evaluate the system [SM07]. The dataset is composed of news headlines which were extracted from news websites (e.g. Google News, CNN). Headlines typically contain several words and are often written with the intention to provoke emotions to attract the attention from the readers. These characteristics of headlines make them suitable to use in emotion recognition and polarity classification tasks. The specific challenge is the small number of words available for the analysis. Although, there is a general intuition that all words can potentially convey affective meaning [SVS06], the coverage of available lexical resources [SV04, ES06, SDS66] falls short for annotating words in headlines. The coverage limitation is actually expected for headlines because of the small number of words. I choose headlines as a dataset because both user utterances in dialogues and headlines are short, so the same difficulty is valid also for user utterances.

The intuition is that the importance or effect of each word for the overall polarity assessment is inversely proportional to the length of a headline. In other words, the contribution of each word to the emotional polarity of a headline increases while the number of words in a headline decreases. Based on this intuition, strictly using the contribution of each word for polarity assessment in a headline is crucial. In the lexical resources for affect or sentiment annotation, there are usually adjectives, verbs, adverbs and common nouns which are useful for emotion or polarity recognition. However, proper nouns (person, location and organization names) in headlines also evoke emotions and positive or negative sentiments for readers. Therefore it has an influence on the polarity of the whole sentence. Even proper nouns may dominate the other sentiment-bearing words which are retrieved from the lexical resources. For example (“*Asia seeks breakthrough on North Korea*”), there are two main sentiment bearing entities, *Asia* and *North Korea*. If I omit these two proper nouns, it is very likely that I will lose the opportunity to recognize the polarity of the sentence. In addition, some of the words or proper nouns may have more influence than others to overall sentiment polarity, in this case, *North Korea* dominates other sentiment bearing words in the headline and it becomes the main influence on the polarity.

The sentiment conveyed by proper nouns evolves by time and current events (e.g. natural or man-made disasters, economic developments, political developments). For example, Japan usually evokes positive sentiments on people but if there is a recent disaster, such as an earthquake or nuclear disaster, it evokes negative sentiments. The proper nouns in sentences can be identified by any standard named entity recognizers [FGM05] which can recognize location, person and organization entities. The real difficulty is deciding the current sentiment polarity of the recognized named entities. I addressed this problem in the system which is discussed in the *Approach* section.

In sentiment polarity recognition, a sentence might contain positive or negative polarity words, but a sentence does not necessarily have the same polarity as a whole. To take into account the interactions between words instead of handling the words independently, I applied dependency parsing and some linguistics rules, such as polarity reversal.

8.2 Related Work

Sentiment analysis has gathered the attention of many research groups from variety of areas, such as affective computing, linguistics and psychology. There is a substantial amount of work done using different approaches, that can be categorized in 2 main categories (i.e. rule-based and statistical approaches). Ruled-based systems usually try to apply linguistic compositionality rules [NPI11] and create highly comprehensive lexicons (i.e. in conjunction with possible compositionality rules) [NPI09] to perform sentiment analysis on text at sentence and phrase level. Machine learning approaches are also frequently applied to the sentiment analysis problem, such as linguistically inspired deep neural networks [SPW⁺13b], graphical models [NIK10] and classical supervised learning techniques [ARS05]. In addition, micro-blogging websites also get the attention of researchers for sentiment analysis [PP10, KWM11]. In these studies, emoticons, as a different modality, are used for sentiment polarity recognition. In this project, I discuss and compare my model to the systems presented at SemEval-2007, where the same corpus of news headlines was used. The main emphasis is on affect of named-entities and prevalent compositionality in headlines.

Five teams participated to SemEval-2007 Task-14: Affective Text, with five systems for valence classification and three systems for emotion labeling. The CLAC system used a knowledge-based, domain-independent, unsupervised approach [AB07]. It uses 3 knowledge sources, a list of sentiment-bearing words, a list of polarity reversal

words, and a set of rules that define the results of combination of sentiment-bearing words with polarity reversal words. The CLAC-NB system uses a Naive Bayes classifier without feature extraction to assess the performance of this basic machine learning technique [AB07]. The UPAR7 system [Cha07] used a rule-based approach, where a list of words for high tech acronyms, celebrities was used in addition to SentiWordNet [ES06] and WordNetAffect [SV04] lexical resources. The SICS used a valence annotation approach based on a word-space model and set of seed words which is based on the idea of creating two points in a high-dimensional word space, one representing positive valence, the other representing negative valence, and projecting each headline into this space [SKE07]. The SWAT system used a supervised methodology by implementing an uni-gram model trained to annotate. Moreover, they added an additional set of 1000 headlines for training [KSW07]. Therefore, the demonstrated systems in SemEval-2007 [SM07], did not consider the contribution of sentiment-bearing named-entities. Since the news headlines very commonly contain proper nouns, there is a good chance to improve the reported results, if the proper names are used in sentiment polarity determination.

8.3 Approach

I used five different approaches for classification after performing the feature extraction. I am using the Stanford dependency parser [DMMM⁺06], the part-of-speech tagger [TKMS03b] and the named-entity recognizer [FGM05] as software packages for feature extraction. I use MPQA Opinion Corpus [WWH05] and General Inquirer [SDS66] as lexical resources to identify sentiment orientation of words. To process the interaction between words, the system uses dependency parsing. The named-entity recognizer is used to recognize locations, organizations (e.g. company names, acronyms) and person names. Then, the syntactic parser is used for word-level anal-

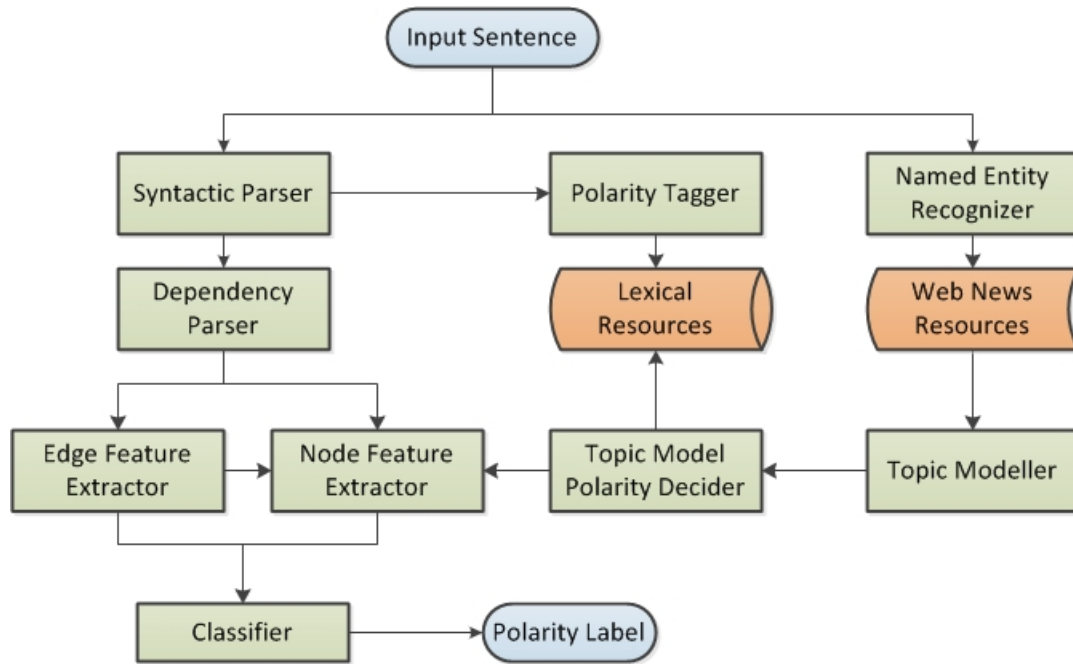


Figure 8.1: System Architecture

ysis. The flow of information among the components of the system is depicted in Figure 8.1.

8.3.1 Named-Entities and Topic Modeling

I observed that most of the news headlines (82% of SemEval DataSet) contain named-entities which influence the overall sentiment polarity of each headline. Therefore, I needed to consider the influence of a named-entity to the overall polarity for an accurate sentiment classification. I am using the Stanford named-entity recognizer to find locations (e.g. Middle East, Gaza), organization names (e.g. Apple, Google, European Union) and person names (e.g. Obama, Madonna), all of which are named-entities. However, the real difficulty is, there is no lexical resource to annotate identified named-entities with the polarity information. To address this problem of finding the sentiment valence for named entities, I performed topic modeling on web news resources (i.e. Google News, CNN). As it is depicted in Figure 8.1, with each identi-

fied named-entity, the system searches on the web to find news articles and gathers top 30 most relevant news articles to the searched named-entity. Then the system performs topic modeling on the news articles retrieved from the web.

Topic modeling provides a simple way to analyze large volumes of unlabeled text. A "topic" consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings. I used the Mallet¹ library for topic modeling.

A topic model with a cluster of 30 words derived from 30 news articles was retrieved from the web. In other words, I represented the recognized named-entity with 30 words to decide the polarity of the named-entity. The cluster of words was tagged using the sentiment polarity lexical resources. The topic model polarity decider component selects the aggregated overall polarity of a topic which actually represents the recognized named-entity. I am using a simple algorithm to decide the overall polarity. If the polarity of a word is positive, it adds 1. If it is negative it subtracts 1, if it is neutral, it does nothing. If the aggregated polarity is greater than 0, the polarity of named-entity is positive. If it is less than 0, the polarity of named-entity is negative.

For example, in Figure 8.2, *Gaza* (city name) is identified as a named-entity. Since it is a proper noun, it does not exist in any of the lexical resources. The system would search the web to retrieve 30 news articles about *Gaza* and perform topic modeling on them to obtain the related word set. The topic modeling component identified "*Israel, air strike, attack, bomb, war, Hamas, terror, refugee, crisis, rocket*" as top 10 words. As previously stated, the system searches for 30 words, all of which are then annotated using the lexical resources. Afterwards, the polarity score for the *Gaza* named-entity is automatically determined as negative using the previous algorithm.

¹<http://mallet.cs.umass.edu/>

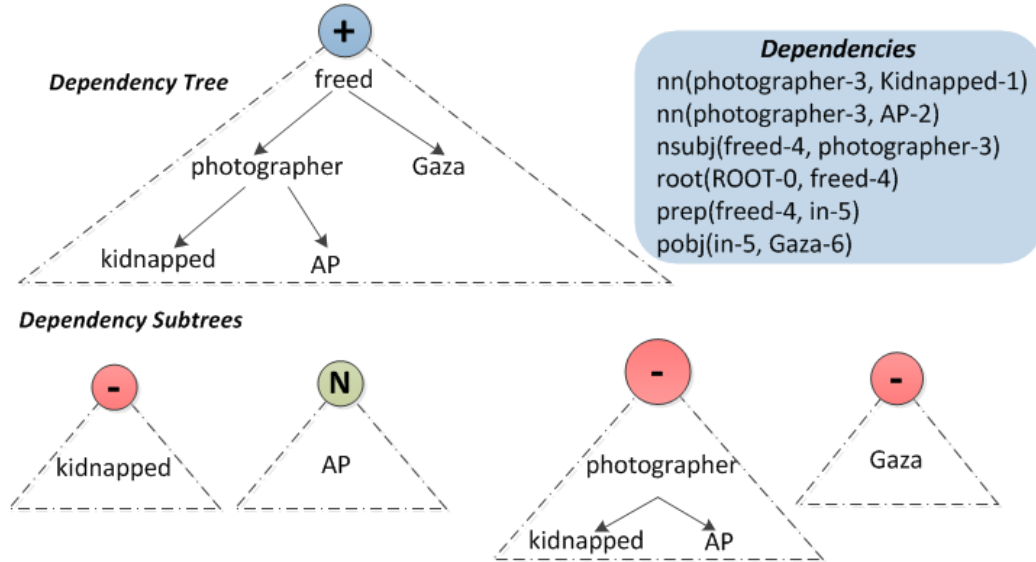


Figure 8.2: Dependencies

8.3.2 Features and Interactions Between Words

The polarity of a whole sentence can not be calculated without considering interactions between words, as it can be seen in Figure 8.2 for the sentence "Kidnapped AP photographer freed in Gaza". There are two negative words (kidnapped, Gaza) and one positive polarity reversal word (freed). If the text is analyzed without considering interactions between words and dependencies, it is very likely that the results might end up with a negative classification. However, as it can be seen in Figure 8.2, even though sub-trees bears negative sentiment, the word "freed" reverses the polarity of the whole sentence.

The system uses two types of features, node features and edge features. Node features represent each node in dependency tree, and edge features represent the interactions between words. Node features include the prior polarity of a word which can be obtained from the lexical resources, polarity reversal, resulting polarity and pos-tag. Edge features represent interaction between sub-trees. For example, in Figure 8.2, kidnapped is negative, AP is neutral which is an organization name,

Node Features	
a	t_i
b	$t_i \& m_i$
c	$t_i \& m_i \& r_i$
d	$t_i \& s_i$
e	$t_i \& c_i$
f	$t_i \& f_i$

Table 8.1: Node Features

photographer is neutral, the negative polarity of *kidnapped* propagates to the subtree where *photographer* is the root. The system not only uses the tree structure for word interactions, but also uses dependencies. For this case, the first dependency, directly shows the relationship between the words *photographer* and *kidnapped*. For the entity *Gaza*, my system uses topic modeling, explained in the *Named-Entities and Topic Modeling* section that yielded a negative polarity. However, the word *freed* reverses the meaning of the whole sentence.

The polarity reversal feature of a node $r_i \in \{0, 1\}$, represents whether or not it reverses the polarity of other words. A polarity reversal word list was prepared so that the property r_i in identified words it is set to 1 otherwise it is 0. The described list was constructed from General Inquirer in the same methodology used in [CC08]. I collected words which belong to either NOTLW or DECREAS categories from General Inquirer (the dictionary contains 121 polarity reversing words). [CC08] categorized polarity reversal words into two distinct categories: function-word negators, such as *not*, and content-word negators, such as *eliminate*. The polarity reversal of a phrase handles only content-word negators, and function-word negators are handled based on the result of Stanford dependency parser which gives the negation (*neg*) relation directly.

In Table 8.1 and Table 8.2, I show the features I used in this project. Features (a)-(f) in Table 8.1 represents node features for the i -th word, which is a node in the

Dependency Features	
A	$t_g \& t_d$
B	$t_g \& t_d \& r_d$
C	$t_g \& t_d \& r_d \& m_d$
D	$t_g \& t_d \& c_g \& c_d \& R$

Table 8.2: Dependency Features

dependency tree. In Table 8.1, t_i denotes polarity of a node, m_i denotes prior polarity of a node. Prior polarity of a node $m_i \in \{+1, 0, -1\}$ is the innate sentiment of a word obtained from the polarity lexical resources. As described before r_i represents polarity reversal word, s_i denotes surface form, c_i denotes coarse-grained part-of-speech (POS) tag, f_i denotes fine-grained POS tag. Features (A)-(D) in Table 8.2 represents dependency features for each dependency governor (head) and its dependent. In Table 8.2, t_g denotes polarity of governor, t_d denotes polarity of dependent, r_d polarity reversal word for dependent, m_d denotes prior polarity as in node features table, c_g denotes coarse-grained pos tag for governor, c_d denotes coarse-grained polarity tag for dependent and R denotes the dependency relationship type (e.g. negation).

After feature extraction including named-entities and interactions between words, I used rule-based and supervised systems to decide the polarity label of a sentence and compare the results.

8.4 Experiments

I conducted the experiments of sentiment classification in SemEval 2007 Affective text news headline corpora.

8.4.1 Data

The dataset consisted of news headlines from major news resources such as New York Times, CNN, and BBC News. The headlines are collected for two main reasons. First, the high load of emotional content, as they describe major national or world-wide events and are written in a style meant to attract the attention of the readers. Second, the structure of headlines was appropriate to make sentence-level annotations of emotions. Two different annotated datasets were made available: one is a development dataset consisting of 250 headlines, and the other with 1,000 headlines. For the experiments I only use valence (positive/negative annotation). The interval for the valence annotations was set to $[-100, 100]$, where 0 represents a neutral headline, -100 represents a highly negative headline, and 100 corresponds to a highly positive headline. Even though, annotations are fine-grained, I used coarse-grained annotations with positive $[50, 100]$, negative $[-100, -50]$ and neutral $(-50, 50)$. Moreover, I used 750 headlines for training and 500 headlines for testing.

It was reported that the test dataset was independently labeled by six annotators, who were instructed to select the appropriate emotions for each headline based on the presence of words or phrases with emotional content, as well as the sentiment polarity invoked by the headline. The agreement evaluations were conducted using the Pearson correlation measure, with an inter-annotator agreement of *78.01*.

8.4.2 Compared Methods

I compared five methods with a different set of features, 3 rule-based methods as used in [NIK10] without using major polarity in training data and 2 supervised classification methods (Naive Bayes and Maximum Entropy) described below. A_i denotes the set of all the ancestor nodes of i -th word in the dependency tree, and $\text{val}(x)$ is defined as:

$$val(x) = \begin{cases} +1, & (x > 1), \\ -1, & (x < 0), \\ 0 & (x = 0). \end{cases}$$

Voting without Polarity Reversal

The polarity of the headline is decided by voting of each node's prior/innate polarity which also includes polarity of each named-entity which is obtained through topic modeling.

$$p = val\left(\sum_{i=1}^n m_i\right) \quad (8.1)$$

Voting with Polarity Reversal

Similar to Voting without polarity reversal, except that the polarities of phrases which have odd numbers of reversal phrases in their ancestors are reversed before voting.

$$p = val\left(\sum_{i=1}^n m_i \prod_{j \in A_i} (-1)^{r_j}\right) \quad (8.2)$$

Rule

The sentiment polarity of a headline is deterministically decided based on rules, by considering the polarities of sub-trees. The polarity of the sub-tree whose ancestor is the i -th word is decided by voting the prior polarity of the i -th word and the polarities of the sub-trees whose ancestor nodes are the modifiers of the i -th word. The polarities of the modifiers are reversed if their governor phrase had a reversal word. The decision rule is applied from bottom to top, the polarity of root node is decided at last.

Method	Acc.	Prec.	Rec.	F1
Voting w/o PR	0.6613	0.4916	0.5378	0.5103
Voting w/ PR	0.6253	0.4553	0.5191	0.4851
Rule	0.644	0.4855	0.6195	0.5394
Naïve Bayes	0.6606	0.4921	0.6461	0.5564
Maximum Entropy	0.7673	0.5849	0.7411	0.6518

Table 8.3: Experiment Results for Sentiment Analysis

$$p = val(m_i + \sum_{j:h_j=i} t_i(-1)^{r_i}) \quad (8.3)$$

Supervised Classification

I have used the Naive Bayes and Maximum Entropy methods for classification using the Mallet machine learning package. It is important to note that I used 750 news headlines as a training data and 500 as testing data.

8.4.3 Experiment Results

I have conducted the experiments using five different approaches. The results are presented in Table 8.3. The table columns from left to right shows, method used to decide polarity of a headline, accuracy, precision, recall and F1 measure. The performance of each method is measured with accuracy and F_β rate as shown in Equation 4 [Rij79].

$$F_\beta = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \quad (8.4)$$

The performance of the rule-based approaches is close to each other, the Naive Bayes approach performed slightly better than the rule-based approaches. The Maximum Entropy classifier achieved better results compared to the other 4 methods.

When I compare the results with the accuracy of the systems presented in SemEval-2007 Affective Text task [SM07], my methodology outperformed the demonstrated systems. The highest reported accuracy and F1 score in SemEval are *0.5510* and *0.4243* respectively. My best performing system's (i.e. maximum-entropy classifier) accuracy and F1 score are *0.7673* and *0.6518* (see Table 8.3) respectively. I believe that the main factor in improvement is taking into account the polarity information of the named-entities by performing topic modeling on news articles. It is easily observable that in news headlines there are many proper names and named entities, and the number of the words in each headline is little. If a system does not use sentiment polarity information for a named-entity, it skips important information for overall polarity of a headline. As I discuss in the *Named-Entities and Topic Modeling* Section, upon running the Stanford named-entity recognizer on the trial data provided in SemEval-2007, I find that 82% percent of the headlines contain at least one named-entity.

8.5 Discussion

In this section, I discussed my approach and experiments to perform sentiment polarity classification on the SemEval-2007 Affective Text data. The system uses some of the available sentiment polarity lexical resources and polarity reversal words for feature extraction. Also I performed topic modeling to decide sentiment polarity of each named-entity. The results show that the consideration of named-entities has a positive effect on sentiment polarity classification. My approaches outperformed the systems presented in SemEval-2007. Although I did not test it, I believe that the system can be used in other domains for sentence level sentiment polarity classification. As a future research, it is possible test the system on different domains, such as user utterances (if any data becomes available) and movies reviews datasets.

CHAPTER 9

Conclusions

This chapter describes summary of my contributions and possible future research directions.

9.1 Summary

In this dissertation, I mainly focused on delivery of brief health interventions via multimodal spoken dialogue systems. In the health dialogue systems area, the mainstream dialogue management mechanism is using rule-based approaches. On the other hand, in spoken dialogue systems area, dialogue management mechanisms are based on reinforcement learning. The domain of the problems is usually tourist information domain in RL-based systems. My contribution has impact to two areas dialog systems and health dialogue. First, I applied first time RL-based techniques to real-world health dialogue. Second I used recent methods that are used in SDS community and avoid mainstream dialogue management approaches that are used in the health domain.

I developed tools to reason, recognize concepts, annotate emotional ratings of words, and analyze sentiments. For the dialogue system, I developed a system to recognize of named entities related with the unhealthy behaviors, it has a crucial importance in language understanding. Recognizing named-entities gives possibility to integrate domain related knowledge to discourse. I designed an ontology that enables to reason about behavioral health concepts.

Discussing issues about health behaviors such as heavy drinking are highly emotional for people to talk about (e.g. shame, discouragement, anger, hopefulness, satisfaction, pride). To enable recognizing sentiments and emotions conveyed through user utterances, I developed lexical resources and a sentiment analysis tool which can

recognize sentiments from a small number of words. As a semantic resource, I created two lexicons and one semantic network. Although these resources can be used in the domain of my research, they can be used as general purposes. I created one lexicon for prior affective ratings (valence, dominance and arousal) of words, which is most comprehensive resource for affective ratings up-to-date. I also created a lexicon for prior sentiment polarities. Both lexicons are generated by using unified semantic network that I created by using existing resources that enable to exploit expert knowledge. I evaluated all resources that I developed against state-of-the-art resources. They yielded significantly better performance.

9.2 Future Directions

The spoken dialogue system can be extended in several directions. In terms of dialogue management, the approach can be tested with different health screening and assessment interviews. Development of a user simulator can be very useful to automate testing. From the corpus that I collected for language modelling, it is possible to create virtual patients. Virtual patients can be used to simulate users. This can provide an advantage in system optimizations. Virtual Patients can also be used to train people for brief intervention delivery skills [LPTK09, KPG⁺07, SHJ⁺06]. In addition, dialogue act level simulators can be created from the dialogue corpus that is collected from real interactions. This corpus contains automatic speech recognition output, system dialog action, semantic value of recognition and reward signal for each dialogue turn. User simulations are important for automatic dialogue strategy learning and the evaluation of dialogue strategies. There are variety of different user simulation methods and it is an active research area [ELP97, GHL05, SWSY06]. This dialog corpus can be used to create simulated users. If statistics becomes available for each question system ask (alcohol related facts), the system might have a better

observation model. For example, if the system has data about relationship between excessive alcohol use and abuse indicators (i. e. drinking related accidents, injury or relationship problems), the system might have observation model based on statistics. It might help to track dialogue state more accurately.

Since the system has an embodied conversational agent interface, emotion recognition based on modalities other than text can improve the interaction. Currently, the non-verbal communication only depends on facial expressions, body gestures can improve it. In addition, the system can benefit from having social dialogue mechanism that aims to establish working alliance and rapport with a user.

I believe that there is still room for improvement for the named-entity recognizer that I developed. The ontology can be expanded, and more conceptual relationships might be created. In the named-entity recognizer, I used WordNet [Mil95c] as a semantic network. USN might be used instead of WordNet because it contains multi-source information in addition to WordNet ontologies. It is possible to extract multi-source information from USN by using graph algorithms as we demonstrated. I believe that potential of multi-source information can be used, if behavioral health ontology can be imported to USN. In addition, to expand the underlying ontology, ontology extraction/learning tools can be used such as OntoLT [BS04].

Sentiment analysis module, USN, named-entity recognizer, and behavioral health ontology can be together used to perform emotion and sentiment recognitions. I developed each system/resource separately, and they are tested separately. I believe together they provide required framework for sentiment and emotion recognition. The connection can be done in following way, sentiment analyzer needs to use on sentiment polarity lexicons and named-entities. Behavioral named-entity recognizer (B-NER) can find named-entities in user utterances, topic modeler module in sentiment analyzer can find sentiment polarity of a named-entity that is found by B-NER.

In addition, as a more comprehensive lexicon than MPQA, automatically generated prior sentiment polarity lexicon (described in Chapter 7) can be used. It is possible to compute affective ratings of named-entities by using affective-ratings lexicon. Since the affective ratings lexicon is very comprehensive, it is possible to compute affective ratings of sentences by modifying lexicons that relies on sentiment recognizer.

BIBLIOGRAPHY

- [AB06] Alina Andreevskaia and Sabine Bergler. Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *EACL*, volume 6, pages 209–215, 2006.
- [AB07] Alina Andreevskaia and Sabine Bergler. Clac and clac-nb: Knowledge-based and corpus-based approaches to sentiment tagging. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 117–120, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- [Abu13] Substance Abuse. Mental health services administration (samhsa). highlights of the 2010 drug abuse warning network (dawn) findings on drug-related emergency department visits. the dawn report. july 2, 2012. samhsa web site, 2013.
- [AFB⁺06] James Allen, George Ferguson, Nate Blaylock, Donna Byron, Nathanael Chambers, Myroslava Dzikovska, Lucian Galescu, and Mary Swift. Chester: towards a personal medication advisor. *J. of Biomedical Informatics*, 39(5):500–513, October 2006.
- [AJDJ04] Mokdad AH, Marks JS, Stroup DF, and Gerberding JL. Actual causes of death in the united states, 2000. *JAMA*, 291(10):1238–1245, 2004.
- [ARS05] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586. Association for Computational Linguistics, 2005.
- [AVSP07] Ruth Aylett, Marco Vala, Pedro Sequeira, and Ana Paiva. FearNot! An Emergent Narrative Approach to Virtual Dramas for Anti-bullying Education. volume LNCS 4871, pages 202–205. Springer-Verlag, Berlin Heidelberg, 2007.
- [BBA07] Hana Boukricha and Christian Becker-Asano. Simulating empathy for the virtual human max. *on Emotion and Computing in conj*, 2007.
- [BBC⁺11] Alan W. Black, Susanne Burger, Alistair Conkie, Helen Hastie, Simon Keizer, Oliver Lemon, Nicolas Merigaud, Gabriel Parent, Gabriel Schubiner, Blaise Thomson, Jason D. Williams, Kai Yu, Steve Young, and

- Maxine Eskenazi. Spoken dialog challenge 2010: comparison of live and control test results. In *Proceedings of the SIGDIAL 2011 Conference*, SIGDIAL '11, pages 2–7, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [BDRG10] Ryan Sjd Baker, Sidney K D’Mello, Ma Mercedes T Rodrigo, and Arthur C Graesser. Better to be frustrated than bored: The incidence, persistence, and impact of learners cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4):223–241, 2010.
- [BFL98] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.
- [BG92] Thomas F Babor and Marcus Grant. Programme on substance abuse : project on identification and management of alcohol-related problems. report on phase ii, an randomized clinical trial of brief interventions in primary health care. 1992.
- [BG06] Timothy Bickmore and T Giorgino. Methodological Review : Health Dialog Systems for Patients and Consumers. *Journal of Biomedical Informatics*, 39(5):465–467, 2006.
- [BG10] Timothy Bickmore and Amanda Gruber. Relational agents in clinical psychiatry. *Harvard review of psychiatry*, 18(2):119–30, March 2010.
- [BGHM⁺08] Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, and Jeff Reynar. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*, page 14, 2008.
- [BGP05] Timothy Bickmore, Amanda Gruber, and Rosalind Picard. Establishing the computer – patient working alliance in automated health behavior change interventions. *Patient Education and Counseling*, 59:21–30, 2005.

- [BH06] Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.
- [BHS⁺11] Ellen E. Bouchery, Henrick J. Harwood, Jeffrey J. Sacks, Carol J. Simon, and Robert D. Brewer. Economic costs of excessive alcohol consumption in the u.s., 2006. *American Journal of Preventive Medicine*, 41(5):516 – 524, 2011.
- [Bic05] T.W. Bickmore. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human*, 12(2):293–327, 2005.
- [BKLO03] J. Bos, E. Klein, O. Lemon, and T. Oka. Dipper: Description and formalisation of an information-state update dialogue system architecture. In *4th SIGdial Workshop on Discourse and Dialogue*, pages 115–124, 2003.
- [BL99] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer, 1999.
- [BPJ09] T.W. Bickmore, L.M. Pfeifer, and B.W. Jack. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 1265–1274. ACM, 2009.
- [BPS⁺10] Timothy W. Bickmore, Kathryn Puskar, Elizabeth a. Schlenk, Laura M. Pfeifer, and Susan M. Sereika. Maintaining reality: Relational agents for antipsychotic medication adherence. *Interacting with Computers*, 22(4):276–288, July 2010.
- [BR09] Dan Bohus and Alexander I. Rudnicky. The ravenclaw dialog management framework: Architecture and systems. *Computer Speech & Language*, 23(3):332 – 361, 2009.
- [BS04] Paul Buitelaar and Michael Sintek. Ontolt version 1.0: Middleware for ontology extraction from text. In *Proceedings of the Demo Session at the International Semantic Web Conference (ISWC)*, 2004.
- [BSS11] Timothy W Bickmore, Daniel Schulman, and Candace L Sidner. A reusable framework for health counseling dialogue systems based on

- a behavioral medicine ontology. *Journal of biomedical informatics*, 44(2):183–97, April 2011.
- [BV12] Yves Bestgen and Nadja Vincze. Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior research methods*, 44(4):998–1006, 2012.
- [BW11] Hana Boukricha and Ipke Wachsmuth. Empathy-Based Emotional Alignment for a Virtual Human: A Three-Step Approach. *KI - Künstliche Intelligenz*, 25(3):195–204, May 2011.
- [BWHG09] Hana Boukricha, Ipke Wachsmuth, A. Hofstatter, and Karl Grammer. Pleasure-arousal-dominance driven facial expression simulation. In *Interaction and Workshops of 3rd International Conference on Affective Computing and Intelligent ACII2009*, pages 1–7, Amsterdam, 2009. IEEE.
- [CB03] Justine Cassell and Timothy Bickmore. Negotiated Collusion : Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modeling and UserAdapted Interaction*, 13(1-2):89–132, 2003.
- [CC08] Yejin Choi and Claire Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 793–801. Association for Computational Linguistics, 2008.
- [CD09] David Coyle and Gavin Doherty. Clinical evaluations and collaborative design: developing new technologies for mental healthcare interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09*, pages 2051–2060, New York, NY, USA, 2009. ACM.
- [CDMS07] David Coyle, Gavin Doherty, Mark Matthews, and John Sharry. Computers in talk-based mental health interventions. *Interact. Comput.*, 19(4):545–562, July 2007.
- [CDPW08] Roberta Catizone, Alexiei Dingli, Hugo Pinto, and Yorick Wilks. Information extraction tools and methods for understanding dialogue in a companion. In *LREC*, 2008.

- [CGP10] Senthilkumar Chandramohan, Matthieu Geist, and Olivier Pietquin. Sparse approximate dynamic programming for dialog management. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL '10*, pages 107–115, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [CGSG04a] Scotty Craig, Arthur Graesser, Jeremiah Sullins, and Barry Gholson. Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of Educational Media*, 29(3):241–250, 2004.
- [CGSG04b] Scotty D. Craig, Arthur C. Graesser, Jeremiah Sullins, and Barry Gholson. Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of Educational Media*, 29:241–250, 2004.
- [Cha07] François-Régis Chaumartin. Upar7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 422–425, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- [CHH12] Erik Cambria, Catherine Havasi, and Amir Hussain. Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *FLAIRS Conference*, pages 202–207, 2012.
- [CLH12] Erik Cambria, Andrew Livingstone, and Amir Hussain. The hourglass of emotions. In *Cognitive behavioural systems*, pages 144–157. Springer, 2012.
- [CSSE+09] Kate B Carey, Lori AJ Scott-Sheldon, Jennifer C Elliott, Jamie R Bolles, and Michael P Carey. Computer-delivered interventions to reduce college student drinking: a meta-analysis. *Addiction*, 104(11):1807–1819, 2009.
- [CVL10] Min Chi, Kurt VanLehn, and Diane Litman. Do micro-level tutorial decisions matter: Applying reinforcement learning to induce pedagogical tutorial tactics. In *Intelligent Tutoring Systems*, pages 224–234. Springer, 2010.
- [DMMM+06] Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

- [dRNC⁺06] Fiorella de Rosis, Nicole Novielli, Valeria Carofiglio, Addolorata Cavalluzzi, and Berardina De Carolis. User modeling and adaptation in health promotion dialogs with an animated character., October 2006.
- [Ekm92] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.
- [ELP97] W. Eckert, E. Levin, and R. Pieraccini. User modeling for spoken dialogue system evaluation. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 80–87, Dec 1997.
- [ES06] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422, 2006.
- [FA⁺98] G. Ferguson, J.F. Allen, et al. Trips: An integrated intelligent problem-solving assistant. In *Proceedings of the National Conference on Artificial Intelligence*, pages 567–573. JOHN WILEY & SONS LTD, 1998.
- [fDCC⁺08] Centers for Disease Control, Prevention (CDC), et al. Alcohol-related disease impact (ardi). *Atlanta, GA: CDC*, 2008.
- [FGM05] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05*, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [FL09] Matthew Frampton and Oliver Lemon. Recent research advances in reinforcement learning in spoken dialogue systems. *Knowledge Engineering Review*, 24(4):375–408, 2009.
- [FQH⁺10] G Ferguson, J Quinn, C Horwitz, M Swift, J Allen, and L Galescu. Towards a Personal Health Management Assistant. *Journal of biomedical informatics*, 43(5 Suppl):S13–6, October 2010.
- [fRE05] Pacific Institute for Research and Evaluation. Drinking in america: Myths, realities, and prevention policy. 2005.

- [GBH09] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.
- [GBSM⁺03] Barbara Gerbert, Steven Berg-Smith, Michelle Mancuso, Nona Caspers, Stephen McPhee, Daniel Null, and Judith Wofsy. Using innovative video doctor technology in primary care to deliver brief smoking and alcohol intervention. *Health promotion practice*, 4(3):249–261, 2003.
- [GD11] Art Graesser and Sidney K. DMello. Theoretical perspectives on affect and deep learning. In Rafael A. Calvo and Sidney K. D’Mello, editors, *New Perspectives on Affect and Learning Technologies*, volume 3 of *Explorations in the Learning Sciences, Instructional Systems and Performance Technologies*, pages 11–21. Springer New York, 2011.
- [GDC⁺08] Arthur C Graesser, Sidney K DMello, Scotty D Craig, Amy Witherpoon, Jeremiah Sullins, Bethany McDaniel, and Barry Gholson. The relationship between affective states and dialog patterns during interactions with autotutor. *Journal of Interactive Learning Research*, 19(2):293–312, 2008.
- [GEKH⁺12] Iryna Gurevych, Judith Ecker-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth. Uby: A large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590. Association for Computational Linguistics, 2012.
- [GGT13] Marco Guerini, Lorenzo Gatti, and Marco Turchi. Sentiment analysis: How to derive prior polarities from sentiwordnet. *arXiv preprint arXiv:1309.5843*, 2013.
- [GHL05] Kallirroi Georgila, James Henderson, and Oliver Lemon. Learning user simulations for information state update dialogue systems. In *INTER-SPEECH*, pages 893–896, 2005.
- [GHL06] Kallirroi Georgila, James Henderson, and Oliver Lemon. User simulation for spoken dialogue systems: Learning and evaluation. In *in Interspeech/ICSLP*, 2006.

- [GOL06] Jonathan Gratch, Anna Okhmatovskaia, and Francois Lamothe. Virtual rapport. *Intelligent Virtual*, 2006.
- [GSS07] Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. *ICWSM*, 7, 2007.
- [GWGF07] Jonathan Gratch, Ning Wang, Jillian Gerten, and Edward Fast. Creating rapport with virtual agents. *Intelligent Virtual Agents*, 2007.
- [GWM10] Kallirroi Georgila, Maria K. Wolters, and Johanna D. Moore. Learning dialogue strategies from older and younger simulated users. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL '10*, pages 103–106, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [GWO07] Jonathan Gratch, Ning Wang, and Anna Okhmatovskaia. Can virtual humans be more engaging than real ones? In *12th International Conference on Human-Computer Interaction*. Springer, 2007.
- [HG00] Kate S Hone and Robert Graham. Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering*, 6(3,4):287–303, 2000.
- [HHEA⁺10] Rachel Humeniuk, S Henry-Edwards, Robert Ali, Vladimir Poznyak, Maristela G Monteiro, World Health Organization, et al. The assist-linked brief intervention for hazardous and harmful substance use: a manual for use in primary care/prepared by r. humeniuk [et al]. 2010.
- [HL04] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.
- [HLG08] James Henderson, Oliver Lemon, and Kallirroi Georgila. Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Computational Linguistics*, 34(4):487–511, 2008.
- [HMG11] Lixing Huang, Louis-philippe Morency, and Jonathan Gratch. Virtual Rapport 2.0. In *International Conference on Intelligent Virtual Agents, Intelligence, Lecture Notes in Artificial Intelligence.*, pages 68–79. Springer-Verlag Berlin Heidelberg, 2011.

- [HP13] Florian Holzschuher and René Peinl. Performance of graph query languages: Comparison of cypher, gremlin and native access in neo4j. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 195–204, New York, NY, USA, 2013. ACM.
- [HSD05] Reid K Hester, Daniel D Squires, and Harold D Delaney. The Drinker’s Check-up: 12-month outcomes of a controlled clinical trial of a stand-alone software program for problem drinkers. *Journal of substance abuse treatment*, 28(2):159–69, March 2005.
- [HSO98] Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305:305–332, 1998.
- [JKM⁺10] Filip Jurcicek, Simon Keizer, François Mairesse, Kai Yu, Steve Young, Srinivasan Janarthanam, Helen Hastie, Xingkun Liu, and Oliver Lemon. D5. 4: Proof-of-concept classic appointment scheduling system (system 2). 2010.
- [JL10] Srinivasan Janarthanam and Oliver Lemon. Adaptive referring expression generation in spoken dialogue systems: evaluation with real users. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 124–131, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [JTY12] Filip Jurčiček, Blaise Thomson, and Steve Young. Reinforcement learning for parameter estimation in statistical spoken dialogue systems. *Computer Speech & Language*, 26(3):168–192, 2012.
- [KFNM04] Holger Knublauch, Ray W. Ferguson, Natalya F. Noy, and Mark A. Musen. The protg owl plugin: An open development environment for semantic web applications. pages 229–243. Springer, 2004.
- [KGWW08a] Sin-hwa Kang, Jonathan Gratch, Ning Wang, and J. Watt. Agreeable people like agreeable virtual humans. In *Intelligent Virtual Agents*, pages 253–261. Springer, 2008.
- [KGWW08b] Sin-hwa Kang, Jonathan Gratch, Ning Wang, and J.H. Watt. Does the contingency of agents’ nonverbal feedback affect users’ social anxiety? In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, number Aamas, pages 120–

127. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

- [KH98] GR Krupka and K Hausman. IsoQuest Inc.: Description of the NetOwl (TM) Extractor System as Used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, pages 1–10, 1998.
- [Klu11] T. Kluwer. I like your shirt - Dialogue Acts for Enabling Social Talk in Conversational Agents. In H. Vihjalmsson, Stefan Kopp, Stacy Marsella, and K R Thorisson, editors, *International Conference on Intelligent Virtual Agents, Intelligence*, volume 6895 of *Lecture Notes in Artificial Intelligence.*, pages 14–27. Springer-Verlag Berlin Heidelberg, 2011.
- [KPG⁺07] Patrick Kenny, Thomas D Parsons, Jonathan Gratch, Anton Leuski, and Albert A Rizzo. Virtual patients for clinical therapist skills training. In *Intelligent Virtual Agents*, pages 197–210. Springer, 2007.
- [KPJD13] Uros Krcadinac, Philippe Pasquier, Jelena Jovanovic, and Vladan Devedzic. Synesketch: An open source library for sentence-based emotion recognition. 2013.
- [KPP⁺12] Catriona M Kennedy, John Powell, Thomas H Payne, John Ainsworth, Alan Boyd, and Iain Buchan. Active assistance technology for health-related behavior change: an interdisciplinary review. *Journal of Medical Internet Research*, 14(3), 2012.
- [KSW07] Phil Katz, Matt Singleton, and Richard Wicentowski. Swat-mp:the semeval-2007 systems for task 5 and task 14. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 308–313, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- [KWM11] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! *ICWSM*, 11:538–541, 2011.
- [LAYR13] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rische. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)*, 4(4):19, 2013.

- [Lem09] Oliver Lemon. User simulations for context-sensitive speech recognition in spoken dialogue systems. *Proceedings of the 12th Conference of the European ...*, (1), 2009.
- [LGHS06] Oliver Lemon, Kallirroi Georgila, James Henderson, and Matthew Stuttle. An isu dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the talk in-car system. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, EACL '06, pages 119–122, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
- [LGRT13] William Li, Jim Glass, Nicholas Roy, and Seth Teller. Probabilistic dialogue modeling for speech-enabled assistive technology. In *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, pages 67–72, Grenoble, France, August 2013. Association for Computational Linguistics.
- [LKW02] Diane Litman, Michael Kearns, and Marilyn Walker. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system satinder singh tavexa@ cs. colorado. edu syntek capital new york, ny 10019. *Journal of Artificial Intelligence Research*, 16:105–133, 2002.
- [LL06] Esther Levin and Alex Levin. Evaluation of spoken dialogue technology for real-time health data collection. *J Med Internet Res*, 8(4):e30, Dec 2006.
- [Lo08] Shao-Kang Lo. The nonverbal communication functions of emoticons in computer-mediated communication. *CyberPsychology & Behavior*, 11(5):595–597, 2008.
- [LP+07] Olivier Lemon, Olivier Pietquin, et al. Machine learning for spoken dialogue systems. In *Proceedings of the European Conference on Speech Communication and Technologies (Interspeech'07)*, pages 2685–2688, 2007.
- [LPE97] Esther Levin, Roberto Pieraccini, and Wieland Eckert. Learning dialogue strategies within the markov decision process framework. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 72–79. IEEE, 1997.

- [LPE98] Esther Levin, Roberto Pieraccini, and Wieland Eckert. Using markov decision process for learning dialogue strategies. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pages 201–204. IEEE, 1998.
- [LPTK09] Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27. Association for Computational Linguistics, 2009.
- [LZS09] Tao Li, Yi Zhang, and Vikas Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 244–252, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
- [Mat80] Joseph D Matarazzo. Behavioral health and behavioral medicine: Frontiers for a new health psychology. *American psychologist*, 35(9):807, 1980.
- [McT02] Michael F. McTear. Spoken dialogue technology: enabling the conversational user interface. *ACM Comput. Surv.*, 34(1):90–169, March 2002.
- [McT11] Michael McTear. Trends, challenges and opportunities in spoken dialogue research. In Wolfgang Minker, Gary Geunbae Lee, Satoshi Nakamura, and Joseph Mariani, editors, *Spoken Dialogue Systems Technology and Design*, pages 135–161. Springer New York, 2011.
- [MDS⁺12] Fabrizio Morbini, David DeVault, Kenji Sagae, Jillian Gerten, Angela Nazarian, and David Traum. FLoReS: a forward looking, reward seeking, dialogue manager. In *4th International Workshop on Spoken Dialogue Systems*, Paris, France, November 2012.
- [MF⁺04] Anne Moyer, John W Finney, et al. Brief interventions for alcohol problems: Factors that facilitate implementation. *Alcohol Research and Health*, 28(1):44, 2004.
- [MFD⁺12] Fabrizio Morbini, Eric Forbell, David DeVault, Kenji Sagae, David R. Traum, and Albert A. Rizzo. A mixed-initiative conversational dialogue

- system for healthcare. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL '12*, pages 137–139, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [MFSV02] Anne Moyer, John W. Finney, Carolyn E. Swearingen, and Pamela Vergun. Brief interventions for alcohol problems: a meta-analytic review of controlled investigations in treatment-seeking and non-treatment-seeking populations. *Addiction*, 97(3):279–292, 2002.
- [MGLT12] Teruhisa Misu, Kallirroi Georgila, Anton Leuski, and David Traum. Reinforcement learning of question-answering dialogue policies for virtual museum guides. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 84–93. Association for Computational Linguistics, 2012.
- [MGM⁺11] Leanne M. Mauriello, N. Simay Gökbayrak, Deborah F. Van Marter, Andrea L. Paiva, and Janice M. Prochaska. An Internet-Based Computer-Tailored Intervention to Promote Responsible Drinking: Findings from a Pilot Test with Employed Adults. *Alcoholism Treatment*, 30(1):1–15, 2011.
- [Mil95a] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.
- [Mil95b] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [Mil95c] George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.
- [MKC10] Sunghwan Mac Kim and Rafael A Calvo. Sentiment analysis in student experiences of learning. In *EDM*, pages 111–120. ERIC, 2010.
- [MKS04] Hans-Michael Muller, Eimear E Kenny, and Paul W Sternberg. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11):e309, 09 2004.
- [ML07] S. McQuiggan and J. Lester. Modeling and evaluating empathy in embodied companion agents. *International Journal of Human-Computer Studies*, 65:348–360, 2007.

- [MNPT02] Bernardo Magnini, Matteo Negri, Roberto Prevete, and Hristo Tanev. A wordnet-based approach to named entities recognition. In *Proceedings of the 2002 workshop on Building and using semantic networks - Volume 11*, SEMANET '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
- [MR02] William R. Miller and Stephen Rollnick. *Motivational Interviewing: Preparing People for Change*, volume 2nd. Guilford Press, New York, 2nd edition, 2002.
- [MR⁺09a] William R Miller, Stephen Rollnick, et al. Ten things that motivational interviewing is not. *Behavioural and Cognitive Psychotherapy*, 37(2):129, 2009.
- [MR09b] William R. Miller and Gary S. Rose. Toward a Theory of Motivational Interviewing. *American Psychologist*, 64(6):527–537, 2009.
- [MRP08] Scott W McQuiggan, Jennifer Robison, and Robert Phillips. Modeling parallel and reactive empathy in virtual agents: An inductive approach. In *on Autonomous agents*, number Aamas, pages 167–174, 2008.
- [MvH04] Deborah L. McGuinness and Frank van Harmelen. OWL web ontology language overview. W3C recommendation, W3C, February 2004.
- [MW02] William R Miller and Paula L Wilbourne. Mesa grande: a methodological analysis of clinical trials of treatments for alcohol use disorders. *Addiction*, 97(3):265–277, 2002.
- [NIA06] NIAAA. Niaaa alcohol alert no. 66: Brief interventions, 2006.
- [NIA07a] NIAAA. *Helping Patients Who Drink Too Much, A Clinician's Guide*. 2007.
- [NIA07b] NIAAA. What colleges need to know now: An update on college drinking research, 2007.
- [NIK10] Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 786–794. Association for Computational Linguistics, 2010.

- [NK11] L. Neuhauser and G.L Kreps. Participatory design and artificial intelligence: Strategies to improve health communication for diverse audiences. Cambridge, MA: AAAI Press, 2011.
- [NP10] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics, 2010.
- [NPI09] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Sentifil: Generating a reliable lexicon for sentiment analysis. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6. IEEE, 2009.
- [NPI11] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Affect analysis model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, 17(1):95–135, 2011.
- [NS07] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January 2007. Publisher: John Benjamins Publishing Co.
- [oHS11] US Dept of Health and Human Services. Alcohol screening and brief intervention for youth: A practitioner’s guide. 2011.
- [Osg57] Charles Egerton Osgood. *The measurement of meaning*. Number 47. University of Illinois press, 1957.
- [Par01] W Parrott. *Emotions in social psychology: Essential readings*. Psychology Press, 2001.
- [Pel09] Catherine Pelachaud. Modelling multimodal expression of emotion in a virtual agent. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1535):3539–48, December 2009.
- [PFB01] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.
- [PGK05] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

- [PHA⁺10] Payam Aghaei Pour, M Sazzad Hussain, Omar AlZoubi, Sidney DMello, and Rafael A Calvo. The impact of system feedback on learners affective and physiological states. In *Intelligent Tutoring Systems*, pages 264–273. Springer, 2010.
- [PI05] Helmut Prendinger and M. Ishizuka. The Empathic Companion - A Character-based Interface that Addresses Users Affective States. *Applied Artificial Intelligence*, 19(3-4):267–286, 2005.
- [PKG] Astrid M Von Der Pütten, Nicole C Krämer, and Jonathan Gratch. Whos there? Can a Virtual Agent Really Elicit Social Presence?
- [PKKM12] Alexandros Papangelis, Nikolaos Kouroupas, Vangelis Karkaletsis, and Fillia Makedon. An adaptive dialogue system with online dialogue policy learning. In *Proceedings of the 7th Hellenic conference on Artificial Intelligence: theories and applications*, SETN’12, pages 323–330, Berlin, Heidelberg, 2012. Springer-Verlag.
- [PL08] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
- [PLL⁺10] Ghislain Putois, France Lannion, Romain Laroche, France Issy-les Molineaux, and Philippe Bretier. Enhanced monitoring tools and online dialogue optimisation merged into a new spoken dialogue system design experience. In *Proceedings of the SIGDIAL 2010 Conference*, pages 185–192, 2010.
- [Plu01] Robert Plutchik. The Nature of Emotions. *American Scientist*, 89(4):344+, 2001.
- [PP08] Tim Paek and Roberto Pieraccini. Automating spoken dialogue management design using machine learning: An industry perspective. *Speech communication*, 50(8):716–729, 2008.
- [PP10] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, 2010.
- [PPM04] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet::similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations ’04, pages 38–41, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

- [PSSJC08] David B Portnoy, Lori a J Scott-Sheldon, Blair T Johnson, and Michael P Carey. Computer-delivered interventions for health promotion and behavioral risk reduction: a meta-analysis of 75 randomized controlled trials, 1988-2007. *Preventive medicine*, 47(1):3–16, July 2008.
- [QGF10] Yu Qiu, Genliang Guan, and Dagan Feng. Improving News Video Annotation with Semantic Context. *2010 International Conference on Digital Image Computing: Techniques and Applications*, pages 214–219, December 2010.
- [Rea05] Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48. Association for Computational Linguistics, 2005.
- [Rij79] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.
- [RKLL11a] Verena Rieser, Simon Keizer, Xingkun Liu, and Oliver Lemon. Adaptive information presentation for spoken dialogue systems: Evaluation with human subjects. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*, 2011.
- [RKLL11b] Verena Rieser, Simon Keizer, Xingkun Liu, and Oliver Lemon. Adaptive information presentation for spoken dialogue systems: evaluation with human subjects. In *Proceedings of the 13th European Workshop on Natural Language Generation, ENLG '11*, pages 102–109, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [RL11] Verena Rieser and Oliver Lemon. Reinforcement Learning for Adaptive Dialogue Systems. 2011.
- [RLL10] Verena Rieser, Oliver Lemon, and Xingkun Liu. Optimising information presentation for spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, pages 1009–1018, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [RN96] B. Reeves and C Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. University of Chicago Press, New York, NY, 1996.

- [RSB⁺11] Heleen Riper, Viola Spek, Brigitte Boon, Barbara Conijn, Jeannet Kramer, Katherina Martin-Abello, and Filip Smit. Effectiveness of e-self-help interventions for curbing adult problem drinking: A meta-analysis. *J Med Internet Res*, 13(2):e42, Jun 2011.
- [RSL01] Charles Rich, Ace L. Sidner, and Neal Lesh. Collagen: Applying collaborative discourse theory to human-computer interaction. *AI Magazine*, 22:15–25, 2001.
- [Rus79] James A Russell. Affective space is bipolar. *Journal of personality and social psychology*, 37(3):345, 1979.
- [SB98] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. Cambridge Univ Press, 1998.
- [SBS11a] Daniel Schulman, Timothy Bickmore, and C.L. Sidner. An Intelligent Conversational Agent for Promoting Long-term Health Behavior Change using Motivational Interviewing. In *2011 AAAI Spring Symposium Series*, pages 61–64, 2011.
- [SBS11b] Daniel Schulman, Timothy W. Bickmore, and Candace L Sidner. An Intelligent Conversational Agent for Promoting Long-Term Health Behavior Change using Motivational Interviewing. In *Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium Series*, pages 61–64. Association for the Advancement of Artificial Intelligence (www.aaai.org), 2011.
- [SCDV⁺98] S. Sutton, R. Cole, J. De Villiers, J. Schalkwyk, P. Vermeulen, M. Maccan, Y. Yan, E. Kaiser, B. Rundle, K. Shobaki, et al. Universal speech tools: The csu toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 3221–3224. Sydney, Australia., 1998.
- [Sch05] Karin Kipper Schuler. Verbnet: A broad-coverage, comprehensive verb lexicon. 2005.
- [SDS66] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer: A computer approach to content analysis. 1966.
- [SFMB07] Horacio Saggion, Adam Funk, Diana Maynard, and Kalina Bontcheva. Ontology-based information extraction for business intelligence. In *Proceedings of the 6th international The semantic web and 2nd Asian con-*

ference on Asian semantic web conference, ISWC'07/ASWC'07, pages 843–856, Berlin, Heidelberg, 2007. Springer-Verlag.

- [SGY05] Jost Schatzmann, Kallirroi Georgila, and Steve Young. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th SIGdial Workshop on DISCOURSE and DIALOGUE*, 2005.
- [SHJ⁺06] Amy Stevens, Jonathan Hernandez, Kyle Johnsen, Robert Dickerson, Andrew Raij, Cyrus Harrison, Meredith DiPietro, Bryan Allen, Richard Ferdig, Sebastian Foti, et al. The use of virtual patients to teach medical students history taking and communication skills. *The American Journal of Surgery*, 191(6):806–811, 2006.
- [SKE07] Magnus Sahlgren, Jussi Karlgren, and Gunnar Eriksson. Sics: Valence annotation based on seeds in word space. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 296–299, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- [SKL⁺00] Satinder Singh, Michael Kearns, Diane J Litman, Marilyn A Walker, et al. Empirical evaluation of a reinforcement learning spoken dialogue system. In *AAAI/IAAI*, pages 645–651, 2000.
- [SLKW02] Satinder Singh, Diane Litman, Michael Kearns, and Marilyn Walker. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *Journal of Artificial Intelligence Research*, 16:105–133, 2002.
- [SM05] Lei Shi and Rada Mihalcea. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *Computational Linguistics and Intelligent Text Processing*, pages 100–111. Springer, 2005.
- [SM07] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics, 2007.
- [SPF⁺07] Richard Saitz, Tibor P Palfai, Naomi Freedner, Michael R Winter, Alexandra Macdonald, John Lu, AL Ozonoff, David L Rosenbloom, and William Dejong. Screening and brief intervention online for college students: the ihealth study. *Alcohol and Alcoholism*, 42(1):28–36, 2007.

- [SPW⁺13a] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.
- [SPW⁺13b] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.
- [SRC⁺05] Karen L Steinberg, Roger A Roffman, Kathleen M Carroll, Bonnie McRee, TF Babor, M Miller, R Kadden, D Duresky, and R Stephens. *Brief counseling for marijuana dependence: a manual for treating adults*. Center for Substance Abuse Treatment, Substance Abuse and Mental Health Services Administration Rockville, MD, 2005.
- [STA⁺10] William Swartout, David Traum, Ron Artstein, Dan Noren, Paul Debevec, Kerry Bronnenkant, Josh Williams, Anton Leuski, Shrikanth Narayanan, Diane Piepol, et al. Ada and grace: Toward realistic and engaging virtual museum guides. In *Intelligent Virtual Agents*, pages 286–300. Springer, 2010.
- [SV04] Carlo Strapparava and Alessandro Valitutti. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086, 2004.
- [SVS06] Carlo Strapparava, Alessandro Valitutti, and Oliviero Stock. The affective weight of lexicon. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 423–426, 2006.
- [SWSY06] Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The Knowledge Engineering Review*, 21(02):97–126, 2006.
- [SY02] Konrad Scheffler and Steve Young. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proceedings of the second international conference on Human Language Technology Research*, pages 12–19. Morgan Kaufmann Publishers Inc., 2002.

- [tH11] M. ter Maat and D. K. J. Heylen. Flipper: An information state component for spoken dialogue systems. In H. Vilhjálmsón, S. Kopp, S. Marsella, and K. Thórisson, editors, *Proceedings of the 10th international conference on Intelligent Virtual Agents (IVA 2011), Reykjavik, Iceland*, volume 6895 of *Lecture Notes in Computer Science*, pages 470–472, Berlin, September 2011. Springer Verlag.
- [THSh⁺11] Markku Turunen, Jaakko Hakulinen, Olov Ståhl, Björn Gambäck, Preben Hansen, Mari C. Rodríguez Gancedo, Raúl Santos de la Cámara, Cameron Smith, Daniel Charlton, and Marc Cavazza. Multimodal and mobile conversational Health and Fitness Companions. *Computer Speech & Language*, 25(2):192–209, April 2011.
- [TKMS03a] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *IN PROCEEDINGS OF HLT-NAACL*, pages 252–259, 2003.
- [TKMS03b] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.
- [TKSDM03] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
- [TL03] David Traum and Staffan Larsson. *The Information State Approach to Dialogue Management*, pages 325–353. Kluwer Academic Publishers, 2003.
- [TY10] Blaise Thomson and Steve Young. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech & Language*, 24(4):562–588, 2010.
- [War91] W. Ward. Understanding spontaneous speech: the phoenix system. In *Proceedings of the Acoustics, Speech, and Signal Processing, 1991*.

- ICASSP-91., 1991 International Conference, ICASSP '91*, pages 365–367, Washington, DC, USA, 1991. IEEE Computer Society.
- [WD01] Joseph B Walther and Kyle P DAddario. The impacts of emoticons on message interpretation in computer-mediated communication. *Social science computer review*, 19(3):324–347, 2001.
- [WG09] Ning Wang and Jonathan Gratch. Rapport and facial expression. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009*, pages 1–6, Amsterdam, September 2009. IEEE.
- [WG10] Ning Wang and Jonathan Gratch. Don't just stare at me.pdf. In *CHI*, pages 1241–1249, Atlanta, GA, USA, 2010.
- [WHCM05] Scott T Walters, Reid K Hester, Emil Chiauzzi, and Elizabeth Miller. Demon rum: High-tech solutions to an age-old problem. *Alcoholism: Clinical and Experimental Research*, 29(2):270–277, 2005.
- [Whi89] Cynthia Whissell. The dictionary of affect in language. *Emotion: Theory, research, and experience*, 4(113-131):94, 1989.
- [Wil02] Walter C Willett. Balancing life-style and genomics research for disease prevention. *Science (New York, N.Y.)*, 296(5568):695–8, April 2002.
- [Wil08] Jason D Williams. The best of both worlds: unifying conventional dialog systems and pomdps. In *INTERSPEECH*, pages 1173–1176, 2008.
- [Wil10] Jason D Williams. Incremental partition recombination for efficient tracking of multiple dialog states. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5382–5385. IEEE, 2010.
- [WK06] Minjuan Wang and Myunghee Kang. Cybergogy for engaged learning: A framework for creating learner engagement through information and communication technology. In *Engaged learning with emerging technologies*, pages 225–253. Springer, 2006.
- [WKB13] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013.

- [WKS⁺10] Angela White, David Kavanagh, Helen Stallman, Britt Klein, Frances Kay-Lambkin, Judy Proudfoot, Judy Drennan, Jason Connor, Amanda Baker, Emily Hines, and Ross Young. Online alcohol interventions: A systematic review. *J Med Internet Res*, 12(5):e62, Dec 2010.
- [WMM12] John S Wodarski, Samuel MacMaster, and Nichole K Miller. The use of computer technology to reduce and prevent college drinking. *Social work in public health*, 27(3):270–282, 2012.
- [WN05] Scott T Walters and Clayton Neighbors. Feedback interventions for college alcohol misuse: What, why and for whom? *Addictive behaviors*, 30(6):1168–1182, 2005.
- [WRL⁺12] Michael Wiegand, Benjamin Roth, Eva Lasarczyk, Stephanie Köser, and Dietrich Klakow. A gold standard for relation extraction in the food domain. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, volume 51, 2012.
- [WWH05] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics, 2005.
- [WY07] Jason D Williams and Steve Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.
- [YALR13] Ugan Yasavur, Reza Amini, Christine Lisetti, and Naphthali Rische. Ontology-based named entity recognizer for behavioral health. In *The Twenty-Sixth International FLAIRS Conference*, 2013.
- [YGK⁺10a] Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174, 2010.
- [YGK⁺10b] Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174, 2010.

- [YGTW13] Steve Young, M Gašić, Blaise Thomson, and JD Williams. Pomdp-based statistical spoken dialog systems: A review. 2013.
- [YLR13] Ugan Yasavur, Christine Lisetti, and Naphthali Rishe. Modeling brief alcohol intervention dialogue with mdps for delivery by ecas. In *Intelligent Virtual Agents*, pages 92–105. Springer, 2013.
- [YN05] Jeonghee Yi and Wayne Niblack. Sentiment mining in webfountain. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 1073–1083. IEEE, 2005.

APPENDIX A

List of Dialog Actions of Spoken Dialog System

Table A.1: Step 1: Dialog Actions

Dialog Action	Utterance
GreetS	Hello, My name is Lola, I am a virtual character created to have conversations with you and teach you about healthy lifestyles. Today, I'm here to talk about drinking to figure out if you have any issues. May I ask you few questions about your drinking?
ReAskGreetingS	I am sorry. I could not understand your answer. To answer my question, you can say, I want to continue or I don't want to continue.
ReAskGreetingU	I could not understand. Can you please repeat your answer?
NotConfirmedGreet	I think I misunderstood. Let me ask again, may I ask you few questions about your drinking?
GreetEndS	Thanks, if you decide to talk about your drinking later, I am always available for you to talk about your drinking. Good bye.
AskQ1S	Do you sometimes drink beer, wine, or other alcoholic beverages?
AskQ1U	Can you briefly talk about your alcohol consumption?
ReAskQ1S	Sorry, I could not understand your answer. Do you sometimes drink beer, wine, or other alcoholic beverages?
ReAskQ1U	Sorry, I could not understand your answer. Can you briefly talk about your alcohol consumption?
NotConfirmedQ1	I am sorry for the misunderstanding, do you sometimes drink beer, wine, or other alcoholic beverages?
Continued on next page	

Table A.1 – continued from previous page

Dialog Action	Utterance
Q1End	Since your told me that you do not consume alcoholic beverages, you do not need to continue with the screening. Thanks for your time.
AskQ2S	How many times in the past year have you had 5 or more drinks in a day?
AskQ2U	Can you approximately estimate the number of times in the last 12 months you had 5 or more drinks in a day?
ReAskQ2S	Sorry, I could not understand your answer. How many times in the past year have you had 5 or more drinks in a day?
ReAskQ2U	Sorry, I could not understand your answer. Can you approximately estimate the number of times in the last 12 months you had 5 or more drinks in a day?
NotConfirmedQ2	I am sorry for the misunderstanding, can you repeat your answer for estimating the number of times in the past year you had 5 or more drinks in a day?
AdviseQ2	Very good, just for your information, I want to mention to you some maximum recommended limits. For healthy men up to age 65, no more than 4 drinks in a day and no more than 14 drinks in a week are recommended as the maximum drinking limits. For healthy women and healthy men over age 65, no more than 3 drinks in a day and no more than 7 drinks in a week are recommended as maximum drinking limits. It would be great for your health to stay within these limits.
Continued on next page	

Table A.1 – continued from previous page

Dialog Action	Utterance
AskQ3S	On average how many days a week do you have an alcoholic drink?
AskQ3U	How frequently do you have an alcoholic drink each week?
ReAskQ3S	Sorry, I could not understand your answer. On average how many days a week do you have an alcoholic drink?
ReAskQ3U	Sorry, I could not understand your answer. How frequently do you have an alcoholic drink each week?
NotConfirmedQ3	Sorry about that, can you repeat your answer for the number of days in a week you have an alcoholic drink?
AskQ4S	On a typical drinking day, how many drinks do you have? Is it more than 4 or 5 drinks?
AskQ4U	On a typical drinking day, how many drinks do you have?
ReAskQ4S	Sorry, I could not understand your answer. On a typical drinking day, how many drinks do you have? Is it more than 4 or 5 drinks?
ReAskQ4U	Sorry, I could not understand your answer. On a typical drinking day, how many drinks do you have?
NotConfirmedQ4	Sorry for asking again, can you repeat your answer about the number of drinks you have on a typical drinking day?
Continued on next page	

Table A.1 – continued from previous page

Dialog Action	Utterance
TransQ4ToSAbuse	Thanks for the information you have provided about your drinking. Let me mention to you some maximum recommended limits. For healthy men up to age 65, no more than 4 drinks in a day and no more than 14 drinks in a week are maximum drinking limits. For healthy women and healthy men over age 65, no more than 3 drinks in a day and no more than 7 drinks in a week are recommended as maximum drinking limits. Next I will try to determine some important aspects of your drinking. Is that okay?

Table A.2: Step 2 Abuse: Dialog Actions

Dialog Action	Utterance
AskQ1S	Have you more than once driven a car or another vehicle while you were drinking? Or after having had too much to drink?
ReAskQ1S	Sorry, I could not understand your answer. Have you more than once driven a car or other vehicle while you were drinking? Or after having had too much to drink?
AskQ1U	Have you gotten into situations while drinking or after drinking that increased your chances of getting hurt, like swimming, using machinery, or walking in a dangerous area or around heavy traffic?

Continued on next page

Table A.2 – continued from previous page

Dialog Action	Utterance
ReAskQ1U	Sorry, I could not understand your answer. Have you or someone else been injured as a result of your drinking in last year?
NotConfirmedQ1	I am sorry for the misunderstanding, can you repeat your answer please? Have you ever taken risks under the influence such as drinking and driving?
AskQ2S	Have you continued to drink even though you knew it was causing you trouble with your family or friends?
ReAskQ2S	Sorry, I could not understand your answer. Have you ever had trouble with your family or friends because of your drinking?
AskQ2U	Is your drinking causing any problems with your family or friends? If yes, can you briefly mention what kind?
ReAskQ2U	Sorry, I could not understand your answer. Is your drinking causing any problems with your family or friends?
NotConfirmedQ2	I am sorry for the misunderstanding, can you please repeat your answer? Have you ever had trouble with your family or friends because of your drinking?
AskQ3S	Have you had a period when your drinking or being sick from drinking often interfered with taking care of your home or family or caused you job troubles or school problems?
ReAskQ3S	Sorry, I could not understand your answer. Did your drinking cause you to fail any of your responsibilities?
Continued on next page	

Table A.2 – continued from previous page

Dialog Action	Utterance
AskQ3U	Have you had a period when your drinking or being sick from drinking?often interfered with your responsibilities? Can you briefly mention?
ReAskQ3U	Sorry, I could not understand your answer. Have you ever failed to do what is expected from you because of your drinking?
NotConfirmedQ3	Sorry, can you repeat your answer. Did your drinking cause you to fail any of your responsibilities?
AskQ4S	Have you been arrested, held at a police station, or had any other legal problems because of your drinking?
ReAskQ4S	Sorry, I could not understand your answer. Have you ever had legal problems such as arrests, DUIs or any other?
AskQ4U	Have you ever had legal problems because of your drinking?
ReAskQ4U	Sorry, I could not understand your answer. Have you ever had legal problems?
NotConfirmedQ4	Sorry for asking again, can you repeat your answer? Did you have any legal problem because of your drinking such as arrests or DUIs ?
TransToDep	Thanks for the information. In the next section I will try to determine some important dependence related aspects of your drinking. Is that okay?

Table A.3: Step 2 Dependence: Dialog Actions

Dialog Action	Utterance
AskQ1S	Have there been times when you ended up drinking more than you meant to? Or kept on drinking for longer than you intended?
ReAskQ1S	Sorry, I could not understand your answer. Have you had times when you ended up drinking more than you meant to? Or kept on drinking for longer than you intended?
AskQ1U	Have there been times when you ended up drinking more than you meant to? Or kept on drinking for longer than you intended? If yes, can you tell me about those times?
ReAskQ1U	Sorry, I could not understand your answer. Have there been times when you ended up drinking more than you meant to? Or kept on drinking for longer than you intended? If yes, can you tell me about those times?
NotConfirmedQ1	I am sorry for the misunderstanding, can you repeat your answer please? Have there been times when you ended up drinking more than you meant to? Or kept on drinking for longer than you intended?
AskQ2S	Have you found that you have to drink much more than you once did to get the effect you want?
ReAskQ2S	Sorry, I could not understand your answer. Do you need to drink a lot more sometimes to get the same effect?
AskQ2U	Have you found that you have to drink much more than you once did to get the effect you want? Can you describe it?
Continued on next page	

Table A.3 – continued from previous page

Dialog Action	Utterance
ReAskQ2U	Sorry, I could not understand your answer. Do you usually have the same effects with same amount of alcohol?
NotConfirmedQ2	I am sorry for the misunderstanding, can you repeat your answer please? Do you usually have same effects with same amount of alcohol?
AskQ3S	Have you more than once wanted to stop or cut down on your drinking? Or tried more than once to stop or cut down but found you couldn't?
ReAskQ3S	Sorry, I could not understand your answer. Have you more than once wanted to stop or cut down on your drinking?
AskQ3U	Have you more than once wanted to stop or cut down on your drinking? Or tried more than once to stop or cut down but found you couldn't? If yes, what happened?
ReAskQ3U	Sorry, I could not understand your answer. When you try to cut down or quit, do you have signs of withdrawal?
NotConfirmedQ3	Sorry, can you repeat your answer. When you try to cut down or quit, do you have signs of withdrawal such as tremors, sweating, nausea or insomnia?
AskQ4S	When the effects of alcohol are wearing off, have you had trouble sleeping? Found yourself shaking? Nervous nauseous restless sweating or with your heart beating fast?
ReAskQ4S	Sorry, I could not understand your answer. Do you keep drinking despite having recurrent physical or psychological problems?
Continued on next page	

Table A.3 – continued from previous page

Dialog Action	Utterance
AskQ4U	When the effects of alcohol are wearing off, have you had trouble sleeping? Did you find yourself shaking? Or nervous, restless, sweating or with your heart beating fast?
ReAskQ4U	Sorry, I could not understand your answer. Do you keep drinking despite having recurrent problems?
NotConfirmedQ4	Sorry for asking again, can you please repeat your answer? Do you keep drinking despite having recurrent physical or psychological problems?
AskQ5S	Have you had a period when you spent a lot of time drinking? Or being sick or getting over the bad aftereffects of drinking?
ReAskQ5S	Sorry, I could not understand your answer. Do you spend a lot of time drinking? For example, do you loose time from work or school due to drinking?
AskQ5U	Have you had a period when you spent a lot of time drinking? Or being sick or getting over the bad aftereffects of drinking? If yes, can you briefly talk about it?
ReAskQ5U	Sorry, I could not understand your answer. Do you spend a lot of time for drinking?
NotConfirmedQ5	Sorry for asking again, can you repeat your answer? Do you spend a lot of time drinking? For example, do you lose time from work or school due to drinking?
Continued on next page	

Table A.3 – continued from previous page

Dialog Action	Utterance
AskQ6S	In order to drink, have you given up or cut down on activities that were important or interesting to you or gave you pleasure such as spending less time with your family and friends because of your drinking?
ReAskQ6S	Sorry, I could not understand your answer. Do you spend less time with your family and friends because of your drinking?
AskQ6U	In order to drink, have you given up or cut down on activities that were important or interesting to you or gave you pleasure?
ReAskQ6U	Sorry, I could not understand your answer. Do you spend less time on other matters such as activities that had been important or pleasurable?
NotConfirmedQ6	Sorry for asking again, can you repeat your answer? Do you spend less time on activities you like such as spending time with your family and friends?
AskQ7S	Have you continued to drink even though you knew it was making you feel depressed or anxious, or causing health problems?
ReAskQ7S	Sorry, I could not understand your answer. Have you continued to drink even though you knew it was making you feel depressed or anxious, or causing a health problem?
AskQ7U	Have you continued to drink even though you knew it was making you feel depressed or anxious? Or causing a health problem. If yes, can you briefly explain?
Continued on next page	

Table A.3 – continued from previous page

Dialog Action	Utterance
ReAskQ7U	Sorry, I could not understand your answer. Have you continued to drink even though you knew it was making you feel depressed or anxious? Or causing a health problem. If yes, can you briefly tell me about it?
NotConfirmedQ7	Sorry for asking again, can you repeat your answer? Have you continued to drink even though you knew it was making you feel depressed or anxious or causing health problems?
TransToFinal	Thanks for the information you provided. I will just take one or two more minute of your time.

Table A.4: Step 3 At Risk: Dialog Actions

Dialog Action	Utterance
Step1	Sometimes you are exceeding medically recommended limits. I strongly recommend that you cut down or quit and I'm willing to help. Is it okay to talk a few more minutes about your drinking?
Step2S	On a scale of 1 to 10, how ready are you to make a change?
Step2U	Are you willing to consider making changes with your drinking?
Continued on next page	

Table A.4 – continued from previous page

Dialog Action	Utterance
Step3N	As I mentioned before, your drinking is more than what is medically safe. Just for your information, men who drink more than 4 standard drinks in a day or more than 14 per week, and for women who drink more than 3 in a day or more than 7 per week, are at increased risk for alcohol related problems. I really want to help you when you feel ready for making change. Is it okay?
Step3Y	Great, Just for your information, men who drink more than 4 standard drinks in a day or more than 14 per week, and women who drink more than 3 in a day or more than 7 per week are at increased risk for alcohol related problems. I recommend you to stick with this limit or abstain. I would also like you to keep track your drinking, and so I will send you a calendar for that if that's fine with you. Finally, I would like to talk with you again 3 months from now. Is it okay?

VITA

UGAN YASAVUR

2009	BSc., Computer Engineering Izmir University of Economics Izmir, Turkey
2012	M.S., Computer Science Florida International University Miami, Florida
2014	Ph.D Computer Science Florida International University Miami, Florida

PUBLICATIONS AND PRESENTATIONS

U. Yasavur, J. Travieso, C. Lisetti (2014). Exploiting Linked Expert Knowledge for Affective and Sentiment Analysis of Words. In *Elsevier Information Processing & Management*. (under review)

U. Yasavur, C. Lisetti (2014). Intelligent Virtual Agents and Spoken Dialog Systems come together to Deliver Health Interventions. In *Springer Journal on Multimodal User Interfaces*. (In Press)

C. L. Lisetti, R. Amini, U. Yasavur (2013). I Can Help You Change! An Empathic Virtual Agent Delivers Behavior Change Health Interventions. *ACM Transactions on Management Information Systems*, Vol. 4, No. 4, Article 19, 2013.

U. Yasavur, C. Lisetti, and N. Rishe (2014). “Statistical Dialog Manager Design Tool for Health Screening and Assessments” In *Proceedings of Fourteenth International Conference on Intelligent Virtual Agents (IVA 2014)*, (Boston, USA, August 2014).

U. Yasavur, J. Travieso, C. Lisetti (2014). Sentiment Analysis Using Dependency Trees and Named-Entities. In *Proceedings of the 27th International FLAIRS Conference*, (Pensacola Beach, USA, May 2014).

R. Amini, C. Lisetti, and U. Yasavur (2014). Emotionally Responsive Virtual Counselor for Behavior-Change Health Interventions. In *Proceedings of the Design Science Research in Information Systems and Technologies (DESRIST 2014)*, (Miami, USA, May 2014).

- U. Yasavur, C. Lisetti, N. Riske (2013). Modeling Brief Alcohol Intervention Dialogue with MDPs for Delivery by ECAs. In *Proceedings of Thirteenth International Conference on Intelligent Virtual Agents (IVA 2013)*, (Edinburgh, UK, August 2013).
- R. Amini, C.L. Lisetti, U. Yasavur, N. Riske (2013). On-Demand Virtual Health Counselor for Delivering Behavior-Change Health Interventions. In *Proceedings of the IEEE International Conference on Healthcare Informatics*, (Philadelphia, USA, September 2013).
- U. Yasavur, R. Amini, C. L. Lisetti (2013). Ontology-based Named Entity Recognizer for Behavioral Health. *Proceedings of the 26th International FLAIRS Conference*, (St. Pete Beach, USA, May 2013).
- R. Amini, U. Yasavur, C. L. Lisetti (2012). HapFACS 1.0: Software/API for Generating FACS-Based Facial Expressions. In *Proceedings of the ACM 3rd International Symposium on Facial Analysis and Animation (FAA '12)*, (Vienna, AUSTRIA, September, 2012).
- U. Yasavur, R. Amini, C. L. Lisetti (2012). User Modeling for Pervasive Alcohol Intervention Systems. In *Proceedings of the Workshop on Recommendation Technologies for Lifestyle Change, In conjunction with the 6th ACM Conference on Recommender Systems (RecSys'12)*, (Dublin, IRELAND, September 2012).
- C. Lisetti, U. Yasavur, C. De Leon, R. Amini, U. Visser, N. Riske. Building On-demand Avatar-based Health Intervention for Behavior Change. In *Proceedings of the 25th International FLAIRS Conference*, (Marco Island, USA, May 2012).
- C. L. Lisetti, U. Yasavur, U. Visser, and N. Riske (2011). Toward Conducting Motivational Interviewing with an On-Demand Clinician Avatar for Tailored Health Behavior Change Interventions. In *Proceedings of the 5th International ICST Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health'11)*, (Dublin, IRELAND, May 2011).