Automatic Learning and Evaluation of User-Centered Objective Functions for Dialogue System Optimisation

Verena Rieser and Oliver Lemon
School of Informatics
University of Edinburgh
Outline

• **Area:** Dialogue strategy design and optimization via data-driven statistical learning

• **Problem:** Modelling “true” User Satisfaction

• **Techniques:** Reinforcement Learning
  • PARADISE regression models of user satisfaction (Walker et al. 1997, 2000)

• **Meta-evaluation:** comparing learned User Satisfaction models across 3 corpora
Dialogue strategy design methods

**Conventional software life cycle**

1. Design guidelines
2. User tests
3. Strategy hand-coding
4. Strategy optimization

**Automatic strategy optimisation**

\[
Q^\pi(s, a) = \sum_{s'} T^a_{ss'} [R^a_{ss'} + \gamma V^\pi(s')]
\]

**Automatic design by optimization function**

\[
 (= \text{“reward”})
\]

Design by `Best practices' (Paek 2007)
Reinforcement Learning

\[ Q^\pi(s, a) = \sum_{s'} T_{ss'}^a [R_{ss'}^a + \gamma V^\pi(s')] \]
Automatic Strategy Optimization using Reinforcement Learning

V. Rieser and O. Lemon: Optimal Multimodal Presentation
Problem: How to guarantee real User Satisfaction?

“Who picked ‘I Can’t Get No Satisfaction’ to be our on-hold music?”

V. Rieser and O. Lemon: Learning Optimal Multimodal Presentation
Research questions

• “Quality assurance” for reward/objective functions: aim is to optimize for real user preferences

• Can we do better than: “Reward = Task Completion – Dialogue Length”? 

• “Bootstrapping”: is a reward function derived from a small Wizard-of-Oz data collection a valid estimate of real user preferences?
The decision/learning problem

Dialogue policy

Information gathering (DM)
  - askAQuestion
  - implicitConf
  - explicitConf
  - presentInfo

Information presentation (NLG)
  - Multimodal
  - Speech only

V. Rieser and O. Lemon: Learning Optimal Multimodal Presentation
Method

• Build simulated learning environment from data collected in a WOZ user study [Rieser & Lemon, ACL’06]

• Train and test RL policy by simulated interaction [Rieser & Lemon: Interspeech’07, JNLE’08]
  • Compare against supervised learning baseline policy (non-optimised policy)

• Test the 2 policies with real users (17 subjects) [Rieser & Lemon, ACL’08]

• Meta-evaluate: compare results from these 3 corpora [Rieser & Lemon, LREC’08] = this paper!
A Wizard-of-Oz experiment

- 5 point Likert scale
PARADISE evaluation framework

[Walker et al. 1997]

$$US_{\text{subjective}} = \alpha \times N(\kappa) - \sum_{i=1}^{n} w_i \times N(C_i)$$

Automatic estimate of subjective **User Satisfaction** (US) from objective dialogue performance measures, using multivariate linear regression, where:

- $\kappa$: task success
- $C_i$: dialogue objective measures (e.g. dialogue length, Word Error Rate, number of confirmations........)
- $w_i$: weights assigned by regression
Use PARADISE reward model to train via Reinforcement Learning

$$US_{\text{subjective}} = \alpha \times N(\kappa) - \sum_{i=1}^{n} w_i \times N(C_i)$$

V. Rieser and O. Lemon: Optimal Multimodal Presentation
Problems/questions:

• Generality of PARADISE models across different systems and user groups (Walker et al. 2000, Paek 2007)?

• Performance of dialogue strategies optimized using PARADISE reward models?
Evaluation of the PARADISE model

- “Model stability”
  a. Build PARADISE model from WOZ data
  b. Build PARADISE model from real user test data
  c. Compare obtained models

- “Model performance” - prediction accuracy
  - Predict unseen events from the same data set
  - Predict unseen events from other data sets

- Performance of dialogue strategies learned with the PARADISE models
Model stability

TaskEase_WOZ = 1.58 + .12 taskCompl
               + .09 mmScore  - .2 dialogueLength

TaskEase_SL = 3.5 + .54 mmScore  - .34 dialogueLength

TaskEase_RL = 3.8 + .49 mmScore  - .36 dialogueLength

- Regression models show the same trends
Model performance: prediction accuracy

- Predicting unseen events in original system
- Predicting unseen events of new systems

- 10-fold cross validation on same data set
- Cross-system evaluations

**Result:** Prediction accuracy is stable across the models and the systems (~16% error)
- The models generalize well
Performance in dialogue optimization

• 17 subjects, 204 dialogues (half SL, half RL)

• The RL policy significantly outperforms the non-optimised (i.e. SL) policy
  • 18 times more reward (p<0.005)

• Users rate the RL policy on average 10% higher (p<.001)

• See our ACL 2008 paper

V. Rieser and O. Lemon: Learning Optimal Multimodal Presentation
Results

• **Overall:** method for design and optimization of dialogue systems (“bootstrapping”)

• **This paper:** method for meta-evaluation of objective/ reward functions

• Despite learning from small amounts of initial WOZ data, a PARADISE-style objective function is a **stable, reliable, and useful** model of real User Satisfaction

• Moving dialogue system design from “art to science” (Sparck-Jones 1996)
Dialogue strategy design methods

Conventional software life cycle

- Design guidelines
- Strategy hand-coding
- User tests

Automatic strategy optimisation

\[ Q^\pi(s, a) = \sum_{s'} T_{ss'}^a [R_{ss'}^a + \gamma V^\pi(s')] \]

- Optimization function
- Strategy optimization
- User tests

Design by `Best practices' (Paek 2007)

Bootstrap from WoZ data

Reward defined via PARADISE
Future work

• User preferences regarding **Natural Language Generation** decisions (see related papers at LONdial 2008)

• Incremental training with improved representations of user preferences

• More data! (e.g. From France Telecom/Orange Labs in the CLASSiC project)

• Further exploration of non-linear reward functions
Thanks for your time. Curious?

• See the TALK project www.talk-project.org (EC FP6)
• See the CLASSiC project (2008-11) “Computational Learning in Adaptive Systems for Spoken Conversation” (FP7 Cognitive Systems)
• www.classic-project.org