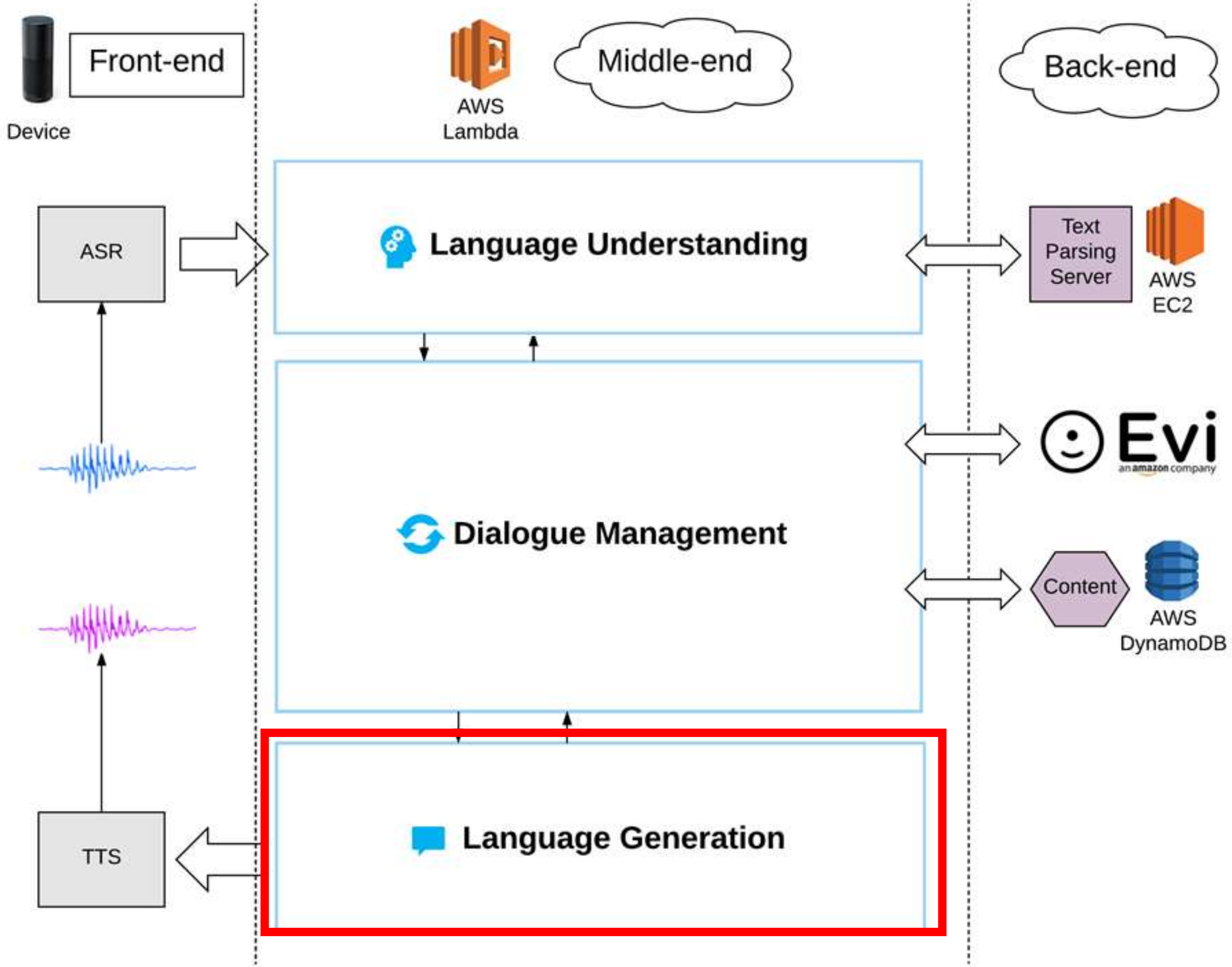
# Natural Language Generation and Dialog System Evaluation

EE596/LING580 -- Conversational Artificial Intelligence

Hao Cheng

University of Washington

# Conv AI System Diagram



# Natural Language Generation

# NLG Approaches

- Template realization
  - use pre-defined templates and fill in arguments
    - ASK_CITY_ORIG: *“What time do you want to leave CITY-ORIG?”*
    - SUGGESTION_TOPIC: *“How about we talk about TOPIC?”*
  - most common in practical systems
- Response retrieval models
  - directly retrieve responses from a large pool
  - active research area, some commercial system uses this approach, e.g., Microsoft XiaoIce
- Response generation models
  - generate the response given the dialog history
  - recent research interest

# IR based model

An new input :

What’s your age ?

**A big conversation corpus**

**A: How old are you**
**B: I am eight**

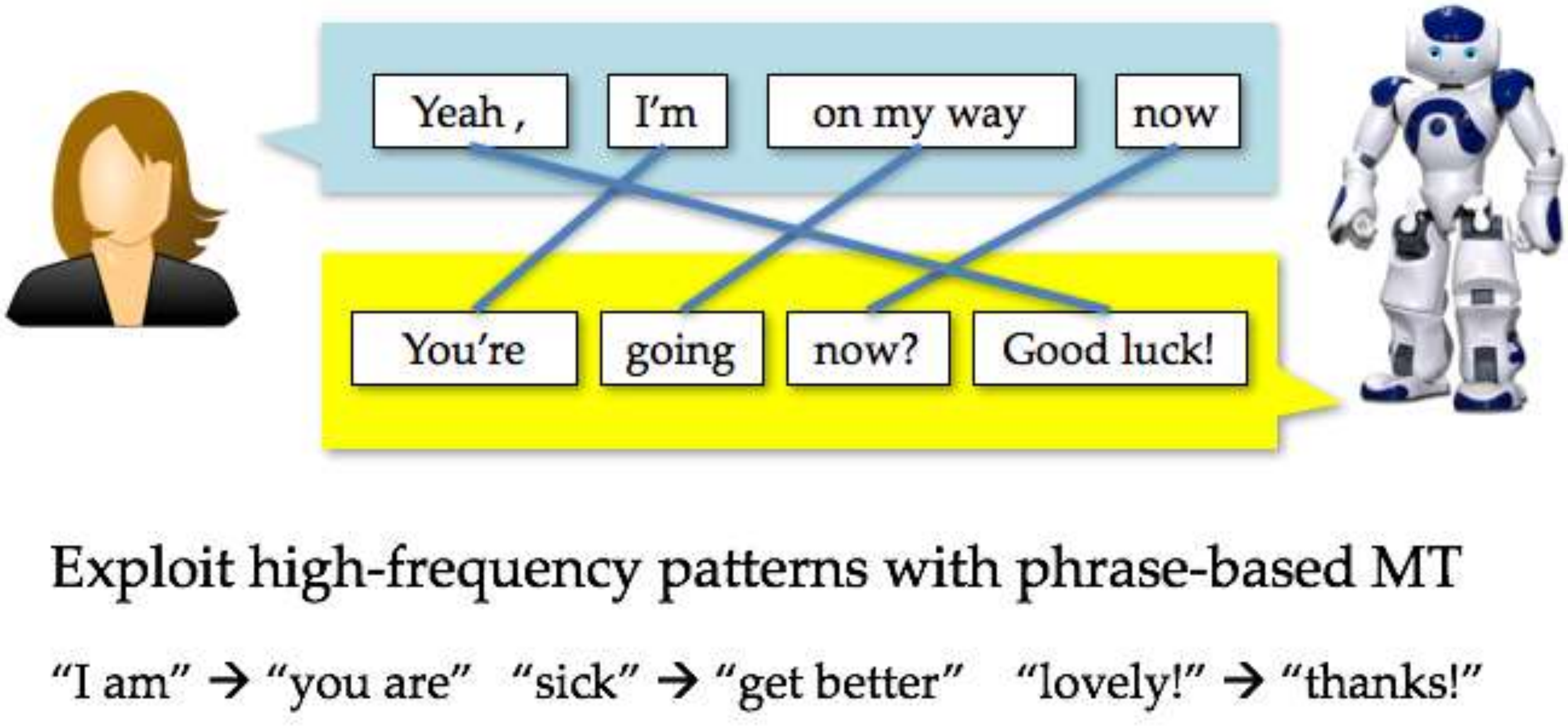
**A: What’s your name ?**
**B: I am john**

**A: How do you like CS224n?**
**B: I cannot hate it more.**

**A: How do you like Jiwei ?**
**B: He’s such a Jerk !!!!!**

# Response Generation as Statistical Machine Translation

(Ritter et al., 2010)




Slide borrowed from Michel Galley

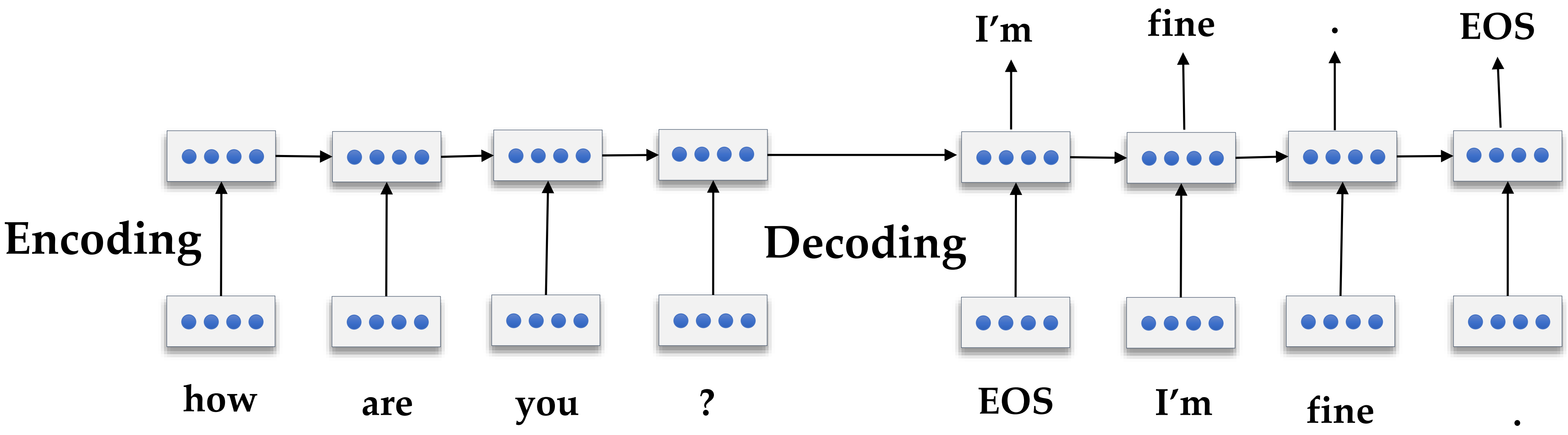
# Seq2Seq Model

(Sutskever et al., 2014; Jean et al., 2014; Luong et al., 2015)

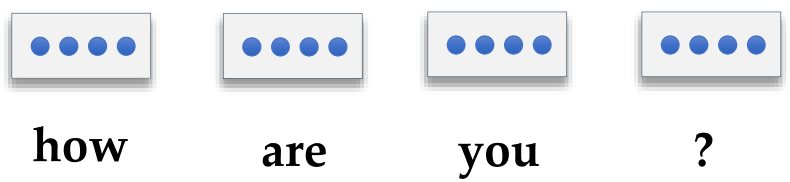
$$\text{Loss} = -\log p(\text{target}|\text{source})$$

**Source : Input Messages**

**Target : Responses**

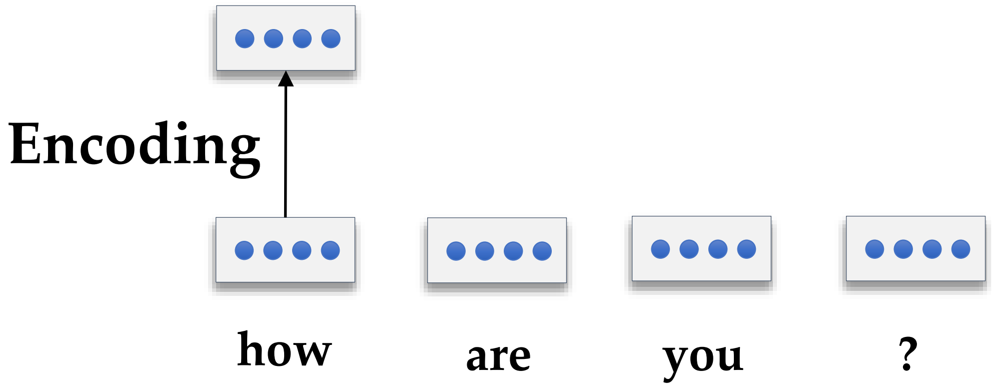


# Seq2Seq Model



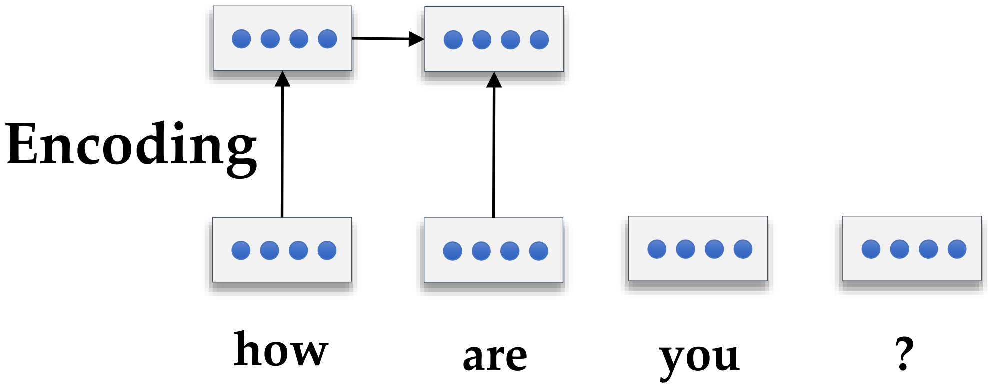

Slide from Jiwei Li, Lecture at CS224S / LINGUIST285 “Spoken Language Processing”

# Seq2Seq Model



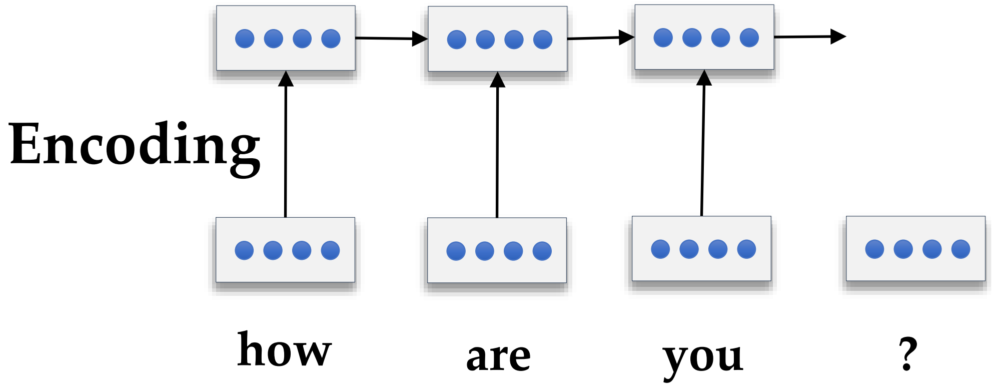

Slide from Jiwei Li, Lecture at CS224S / LINGUIST285 “Spoken Language Processing”

# Seq2Seq Model



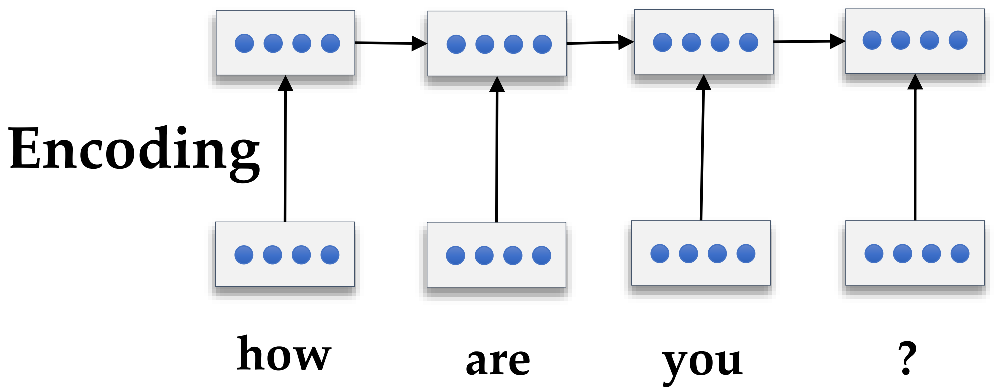

Slide from Jiwei Li, Lecture at CS224S / LINGUIST285 “Spoken Language Processing”

# Seq2Seq Model



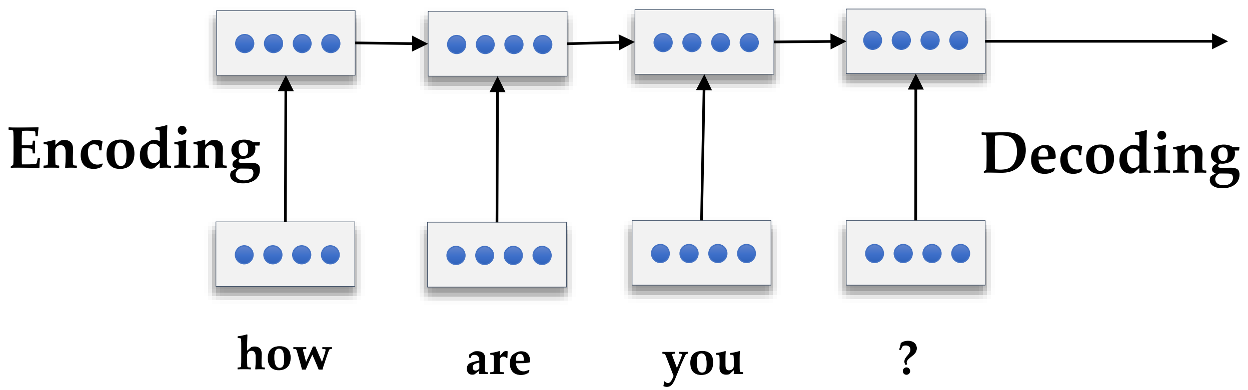

Slide from Jiwei Li, Lecture at CS224S / LINGUIST285 "Spoken Language Processing"

# Seq2Seq Model



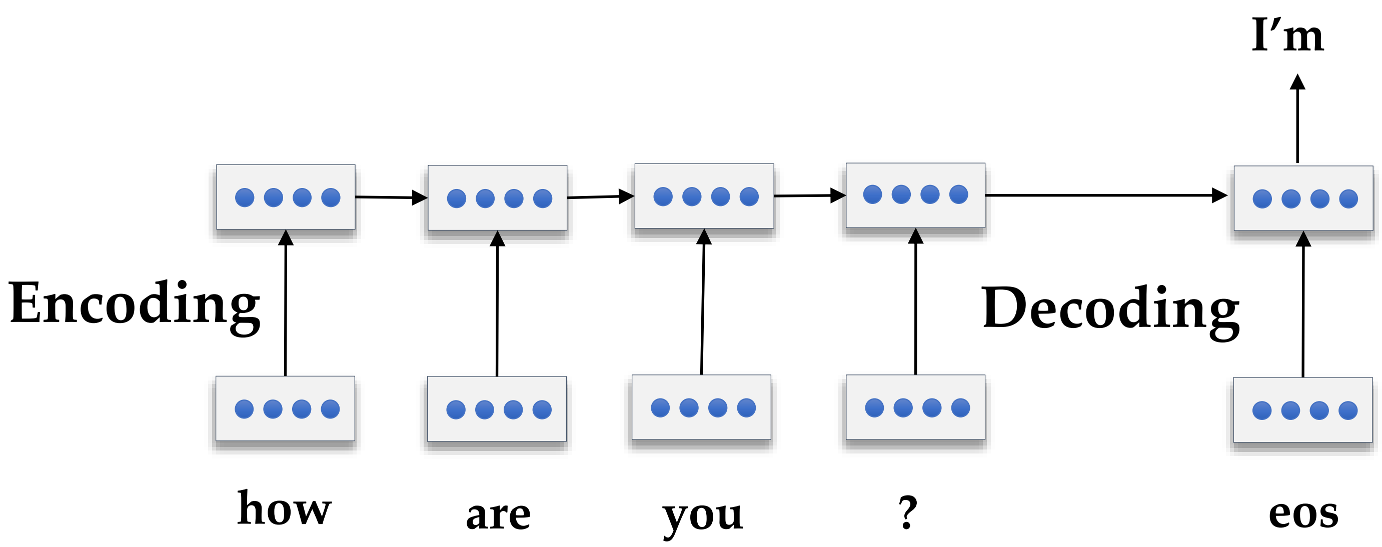

Slide from Jiwei Li, Lecture at CS224S / LINGUIST285 "Spoken Language Processing"

# Seq2Seq Model



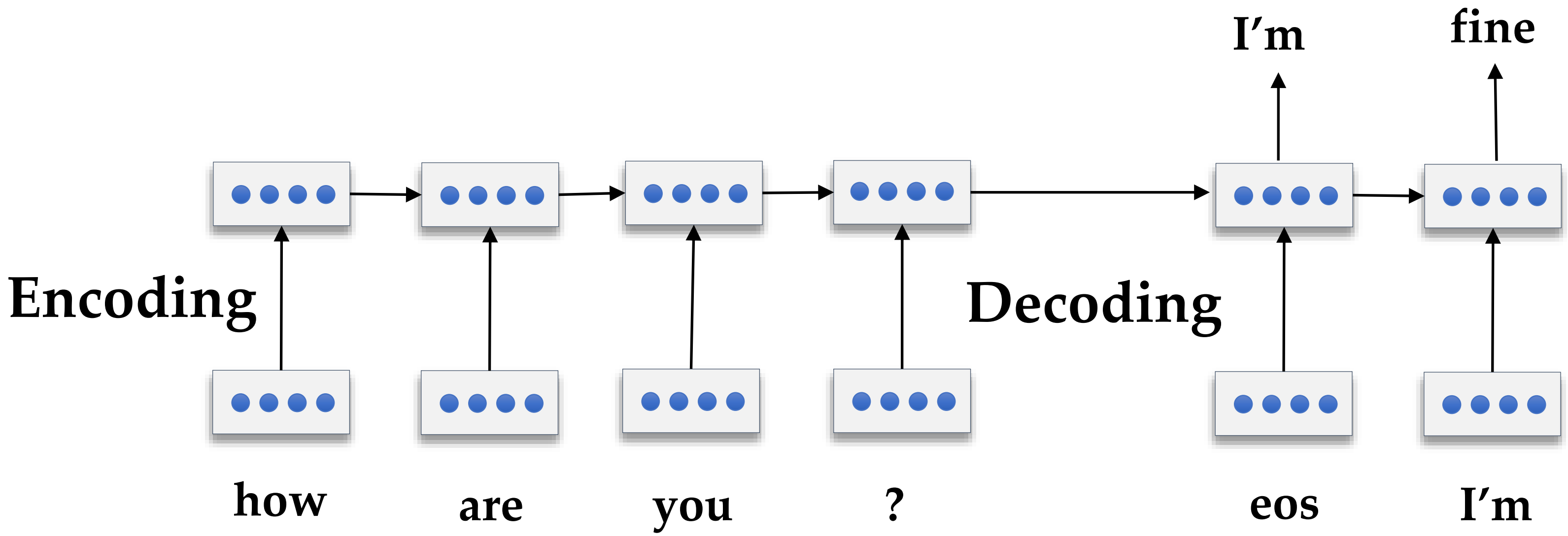

Slide from Jiwei Li, Lecture at CS224S / LINGUIST285 "Spoken Language Processing"

# Seq2Seq Model



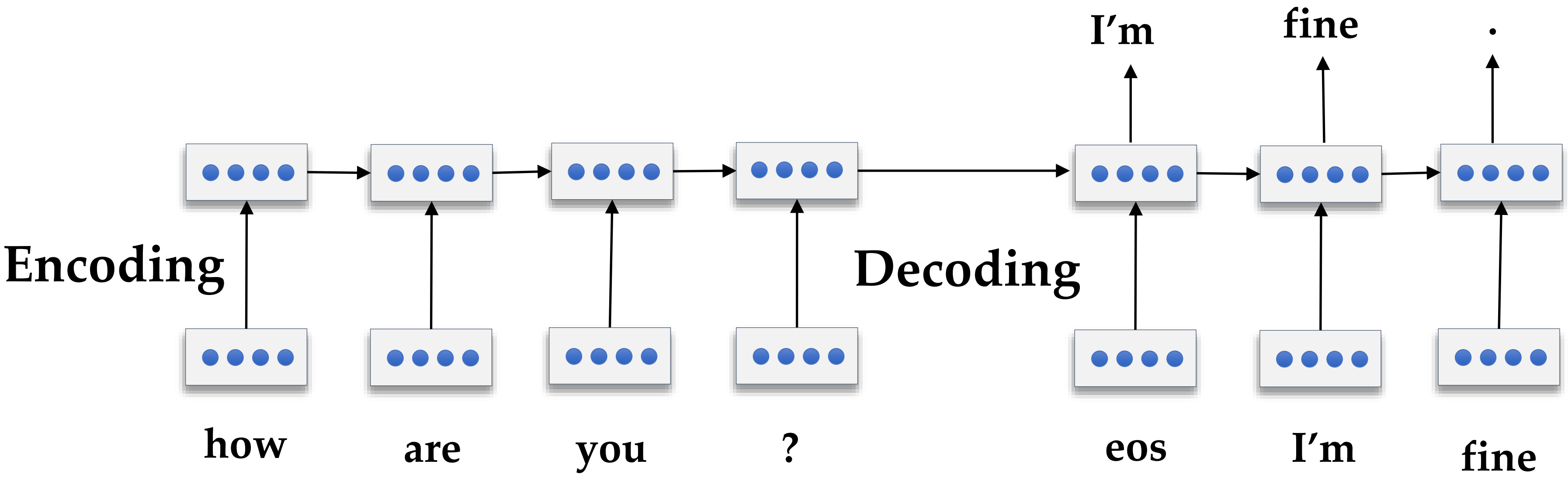

Slide from Jiwei Li, Lecture at CS224S / LINGUIST285 "Spoken Language Processing"

# Seq2Seq Model



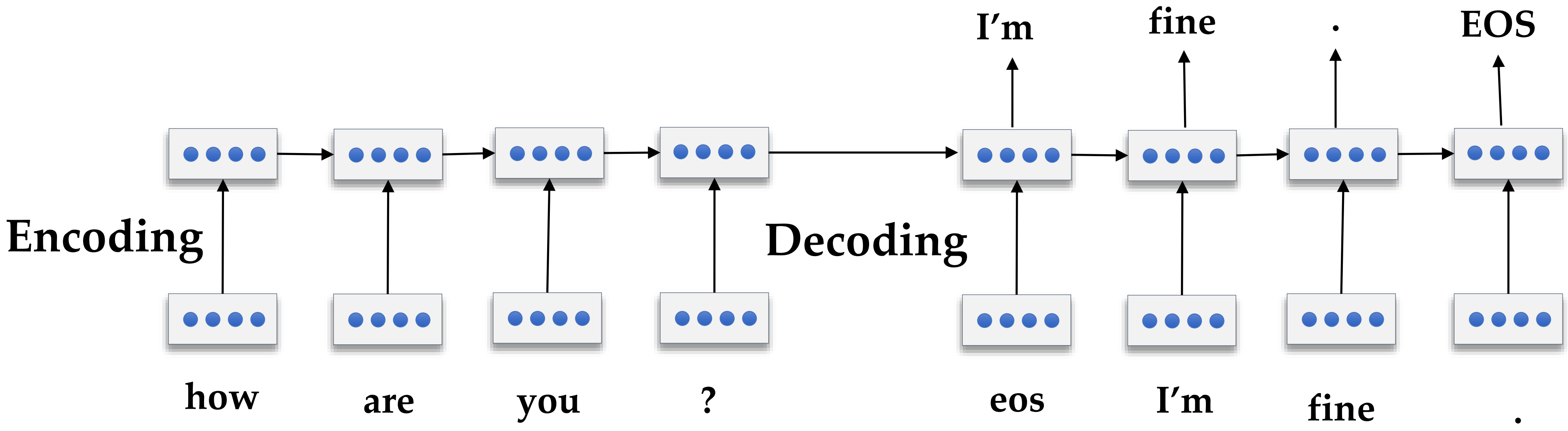

Slide from Jiwei Li, Lecture at CS224S / LINGUIST285 "Spoken Language Processing"

# Seq2Seq Model




Slide from Jiwei Li, Lecture at CS224S / LINGUIST285 "Spoken Language Processing"

# Seq2Seq Model




Slide from Jiwei Li, Lecture at CS224S / LINGUIST285 "Spoken Language Processing"

# Sample Results from Google's paper

| Input | Response |
|---|---|
| what is moral ? | what empowered humanity |
| what is immoral ? | the fact that you have a child . |
| what is the purpose of existence ? | to find out what happens when we get to the planet earth . |
| what do you think about bill gates ? | He's a good man |

Slide from Jiwei Li, Lecture at CS224S / LINGUIST285 "Spoken Language Processing"

# Dialog System Evaluation

# Historical Overview

- In U.S., started around late 1980s by ARPA/DARPA in competitive projects to assess speech technology
  - **A**irline **T**ravel **I**nformation **S**ystem (**ATIS**) project (Price, 1990)
    - speech recognizers for full sentences and read speech
  - **Switchboard** corpus (Jurafsky et al., 1997)
    - collection and annotation of natural telephone conversations
  - **Communicato**r project (Walker et al., 2002)
    - construction and evaluation of spoken dialog systems

# Historical Overview

- In Europe, formulation of standards via collaborative projects
  - **E**xpert **A**dvisory **G**roup on **L**anguage **E**ngineering **S**tandards (**EAGLES**) project (King et al., 1996)
    - a thorough overview of systems and techniques in language engineering
  - **S**peech **R**ecognizer **Q**uality **A**ssessment in **L**anguage **E**ngineering (**SQALE**) project (Young et al., 1997)
    - assessment of large vocabulary, continuous speech recognition systems in a multilingual environment
  - **DISC** project (Bernsen and Dybkjaer, 1997, 2000, 2002)
    - best practices for development and evaluation in dialogue engineering
  - **C**ollaboration in **L**anguage and **S**peech **S**cience and technology (**CLASS**) project (Jacquemin et al., 2000)
    - assessment of speech and language technology with collaboration between EU and US

# Current Industry Practice

- Dialog system evaluation is a standard part of the development cycle
- Extensive testing with real users in real situations is usually done only in companies and industrial environments
- Guidelines and recommendations of best practices are provided in large-scale industrial standardization work
  - International Organization for Standardization (ISO)
  - World Wide Web Consortium (W3C)
- General methodology and metrics are still research issues

# Current Research Efforts

- Shared resources that facilitate prototyping and comparisons
  - Infrastructure: Alexa Skill Kits, Amazon Lex, Facebook ParlAI, Google's DialogFlow, Microsoft BotFramework & LUIS, Rasa, …
  - Corpora: DSTC, Ubuntu chat corpus, DailyDialog, … (see a comprehensive list at https://breakend.github.io/DialogDatasets/)
- Competitions
  - Amazon Alexa Prize, ConvAI challenges, DSTC, …
- Automatic evaluation and user simulations
  - enable quick assessment of design ideas without resource-consuming corpus collection and user studies
- Address new evaluation challenges brought by development of more complex and advanced dialog systems
  - multimodality, conversational capability, naturalness, …

# Basic Concepts

# Evaluation Conditions

- Real-life conditions (field testing)
  - Observations of the users using the system as part of their normal activities in actual situations
  - (Generally) providing the best conditions for collecting data
  - Costly due to the complexity of the evaluation setup
- Controlled conditions (laboratory)
  - Tests take place in the development environment or in a particular usability laboratory
  - (Often) the preferred form of evaluation, but …

# Issues in Controlled Conditions

- Do not necessarily reflect the difficult conditions in which the system would be used in reality
  - Task descriptions and user requirements may be unrepresentative of some situations that occur in authentic usage contexts
- Differences between recruited subjects and real users (Ai et al. 2007)
  - subjects talk significantly longer than users
  - subjects are more passive than users and give more yes/no answers
  - task completion rate is higher for subjects than users

# Theoretical vs. Empirical Setups

- More theoretically oriented setups
  - verify the consistency of a certain model
  - assess predictions that the model makes about the domain
- Less theoretically oriented setups (more empirical)
  - collect data on the basis of which empirical models can be compared and elaborated
- Both approaches can be combined with evaluations in laboratory or real usage conditions

# Types of Evaluation

- Functional evaluation
  - pin down if the system fulfills the requirements set for its development
- Performance evaluation
  - assess the system’s efficiency and robustness in achieving the task goals
- Usability evaluation
  - measure the user’s subjective views & satisfaction
- Quality evaluation
  - measure extra value (e.g., trust) brought to the user through interactions
- Reusability evaluation
  - assess the ease of maintain and upgrade the system

# Evaluation Measures

- Qualitative evaluation: form a conceptual model of the system
  - what the system does?
  - why errors or misunderstandings occur?
  - which parts of the system need to be altered?
- Quantitative evaluation: obtain quantifiable information about the system
  - e.g., task completion, dialog success, …
  - descriptions of the evaluation can still be subjective, the quantified metrics are regarded as objective
  - the objectiveness of a metric can be measured by the inter-annotator agreement (e.g., the Cohen's kappa coefficient you computed in Lab 3)

# Evaluation Measures

- Task-oriented systems
  - Efficiency: length of the dialog, mean user & system response time, the number of help requests/barge-ins/repair utterances, correction rate, timeouts, …
  - Effectiveness: number of completed tasks and subtasks, transaction success, …
  - Usability: user's opinions, attitudes, and perceptions of the system through questionnaires and personal interviews

# Evaluation Measures (Cont.)

- Non-task-oriented system & open-domain chatbots
  - human ratings from either experts or crowdsourced workers
    - annotate system responses based on coherence and appropriateness
    - user self-reported ratings (turn-level and conversation-level)
    - expensive to collect
  - reference-based evaluation
    - widely used in recent neural response generation models
    - measure the similarity between individual system responses and their corresponding reference responses, e.g., perplexity, BLUE, METEOR, ROUGE
    - weak correlation with human ratings at turn-level (Liu et al. 2016)
    - does not account for the fact that responses with completely different meanings can be equally acceptable

# Evaluation Measures (Cont.)

- Non-task-oriented system & open-domain chatbots
  - model-based evaluation
    - supervised models to predict human ratings of candidate bot responses
    - need to collect a large amount of data
    - may be generalize to other domains / datasets
  - reward functions in reinforcement learning
    - can be treated as an evaluation metric
    - mostly hand-crafted (e.g., scores measuring the ease of answering, information flow, and semantic coherence)
    - can also be learned from data (similar to the model-based evaluation)

# PARADISE Evaluation Framework for Task-Oriented Systems

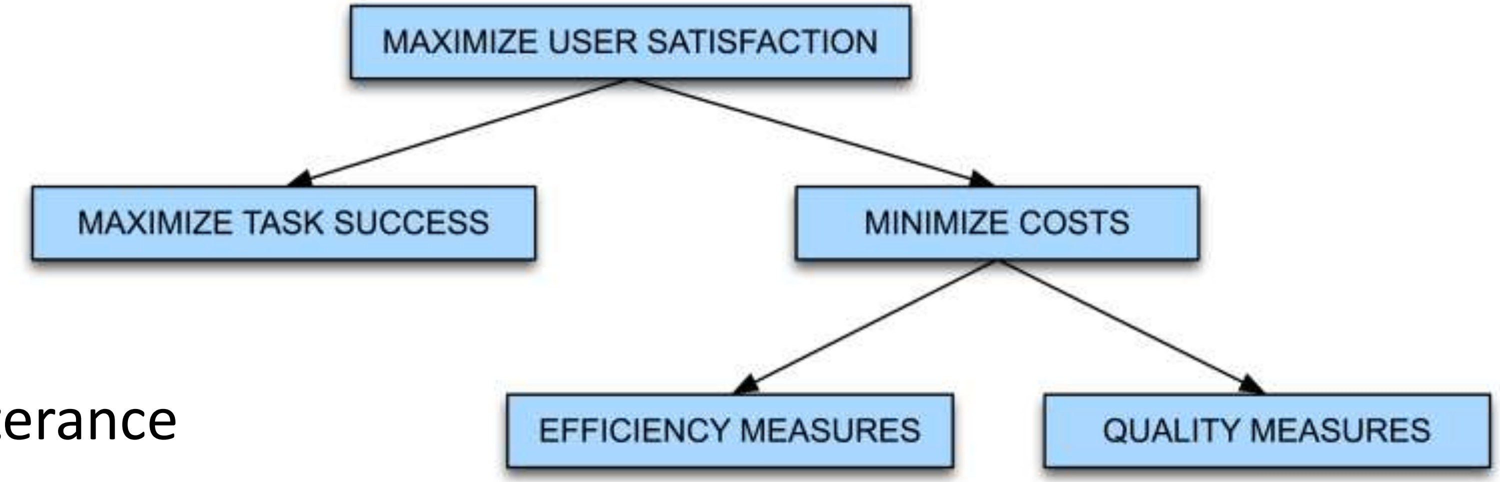
# PARADISE (Walker et al. 2000)

- **PARA**digm for **D**ialogue **S**ystem **E**valuation
  - measure the system's performance with the help of features related to task success and task costs
  - widely used in the literature for task-oriented systems
- Approach
  - learn a linear regression model to estimate conversation-level user satisfaction using a set of features
  - (hopefully) the model learns to
    - maximize a subset of features representing task success
    - minimize a subset of features representing task cost

# Experimental Procedures

- Users given specified tasks & spoken dialogs recorded
- Cost factors, states, dialog acts automatically logged
- ASR accuracy, barge-in hand-labeled
- Users specify task solution & complete satisfaction surveys
- Learn a linear regression model to estimate user satisfaction as a function of task success and costs
  - involves feature selection
  - test for significant predictive features

# Features




**Task success**

- % of subtasks completed
- Correctness of each system utterance
- Correctness of total solution
- Users' perception of task completion

**Efficiency cost**

- Total elapsed time in seconds or turns
- Number of queries
- Turn correction ratio

**Quality cost**

- ASR accuracy
- # of ASR rejection prompts
- # of times user had to barge-in
- # of time-out prompts
- Inappropriateness of system response

# An Example Performance Function

$$\text{Performance} = 0.45N\,(\text{Comp}) + 0.35N\,(\text{MR}) - 0.42N\,(\text{BI})$$

- COMP: User perception of task completion (success)
- MR: Mean (concept) recognition accuracy (cost)
- BI: barge-ins (cost)

- Allows comparing systems as long as the same metrics are used.
- If the system do not exhibit the same interaction possibilities (e.g., does not allow barge-ins), a straightforward comparison is not possible.

# Issues in PARADISE

- High cost for deriving the performance function
  - requires elaborate data collection including the setting up of user tests and the annotation and analysis of the collected data
  - may be practically impossible to collect enough representative dialogs
- Linear superposition of interaction parameters seems too simplistic for such a complex task
  - the correlations between user judgments and interaction parameters remain weak (Moller 2009)
- It is not clear if the predictions are dependent on the particular system.
  - prediction power significantly reduced if the users of the system are changed from novices to experts (Walker 2000 et al.)
  - extrapolation from one system to another significantly reduces prediction power (Moller 2005)

# Current Evaluation Approaches for Socialbots

# Alexa Prize Socialbots Evaluation

- Evaluated primarily by Alexa users who give a rating upon finishing their conversations with a socialbot
- University teams mostly use user ratings for assessing the system performance and perform A/B testing for system diagnosis
- Besides the conversation-level user ratings, teams also use conversation duration and number of turns to assess conversation quality

# Proxy Metrics for User Ratings

- Number of total turns
  - several teams find it positively correlates with user ratings, although the correlation is relatively weak
- User sentiment
  - slightly correlated with user ratings in several studies
- Percentage of user turns with positive/negative reactions (identified by pre-defined key phrases and automatically derived sentiment polarity)
  - % positive user turns: positively correlated with user ratings, although the correlation is as low as the number of total turns
  - % negative user turns: much lower correlation

# User Characteristics vs. User Ratings

- User’s mood affects their ratings (Larionov et al., 2018)
  - users classified as in a great mood rate conversations on average 1.4 point higher than those classified as unhappy.
- Users who curse more tend to rate the conversation lower than normal users (Ji et al., 2017)
- Frequent users who have had at least two conversations with a particular socialbot give lower ratings than general users (Venkatesh et al., 2017)
- User personality traits are correlated with user ratings (Fang et al. 2018)
  - users that are more extraverted, agreeable, or open to experience tend to rate the conversation higher

# User Ratings Prediction

- Deep neural networks and ensemble models (Venkateshetal., 2017)
  - n-grams of user-bot turns, token overlap between user utterance and socialbot response, conversation duration, number of turns, and mean response time
- an ensemble of linear regression models (Serban et al. 2017)
  - dialog length, sentiment, genericness, length, confusion, appropriateness
  - used for rewards in reinforcement learning