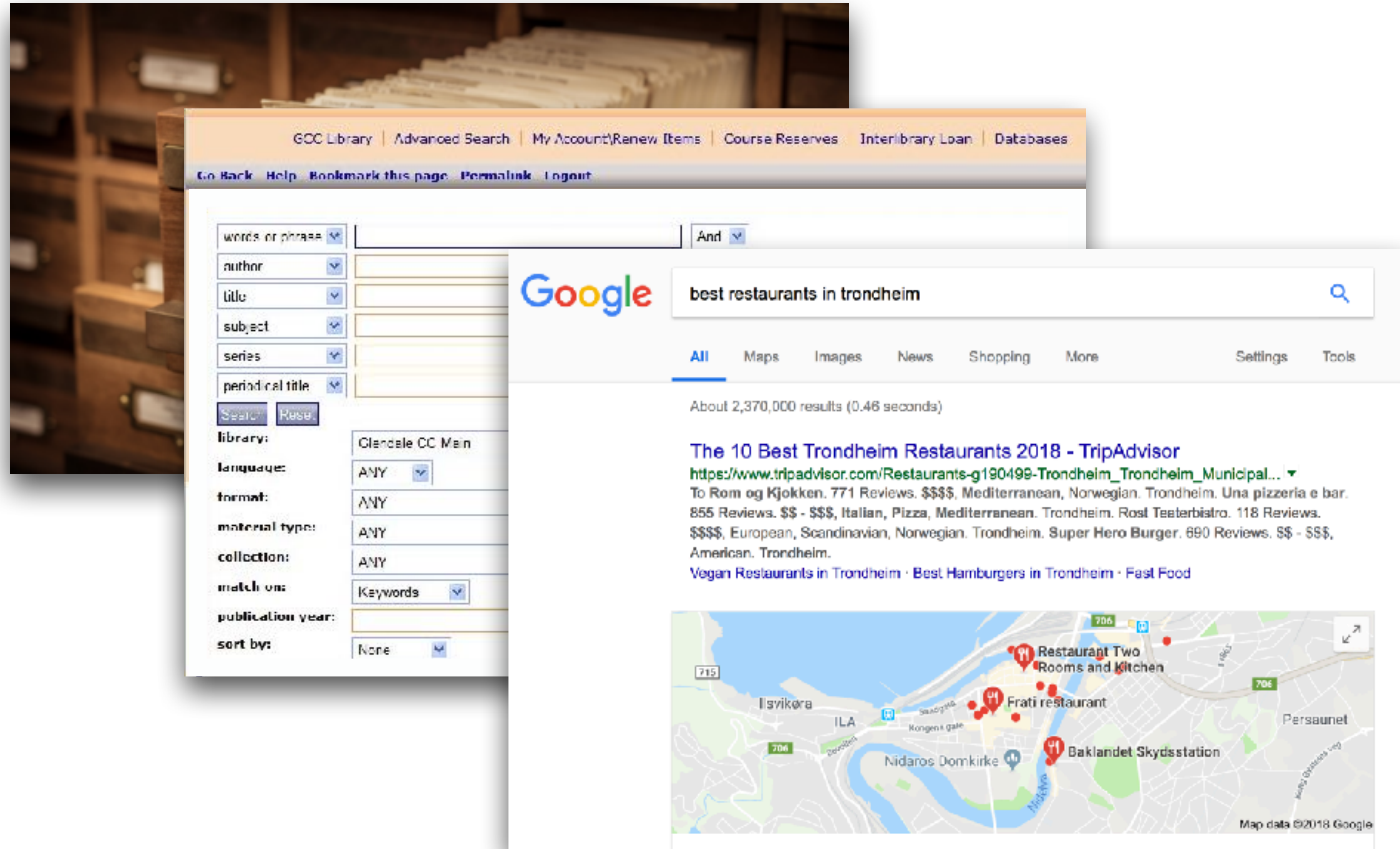
CONVERSATIONAL IR

FAEGHEH HASIBI

MARCH 12, 2018

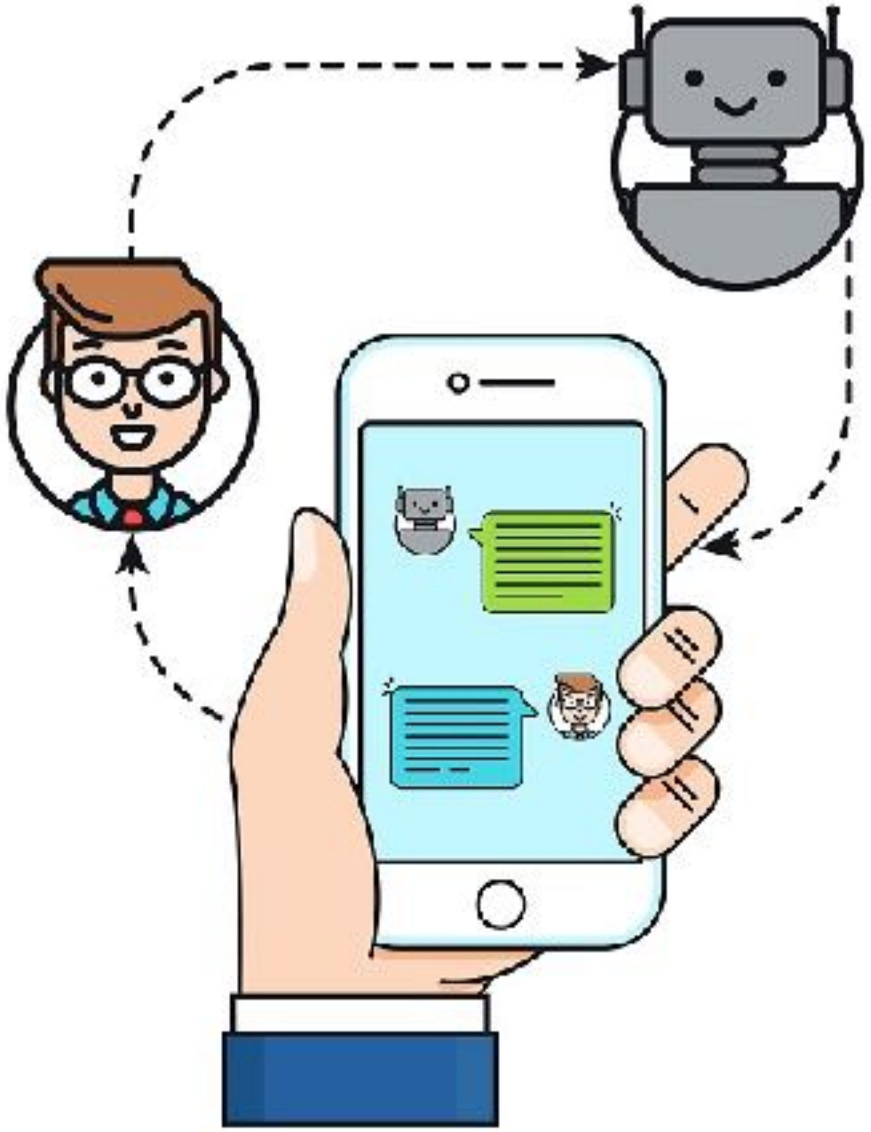
NTNU
Norwegian University of Science and Technology

IR - Information Retrieval



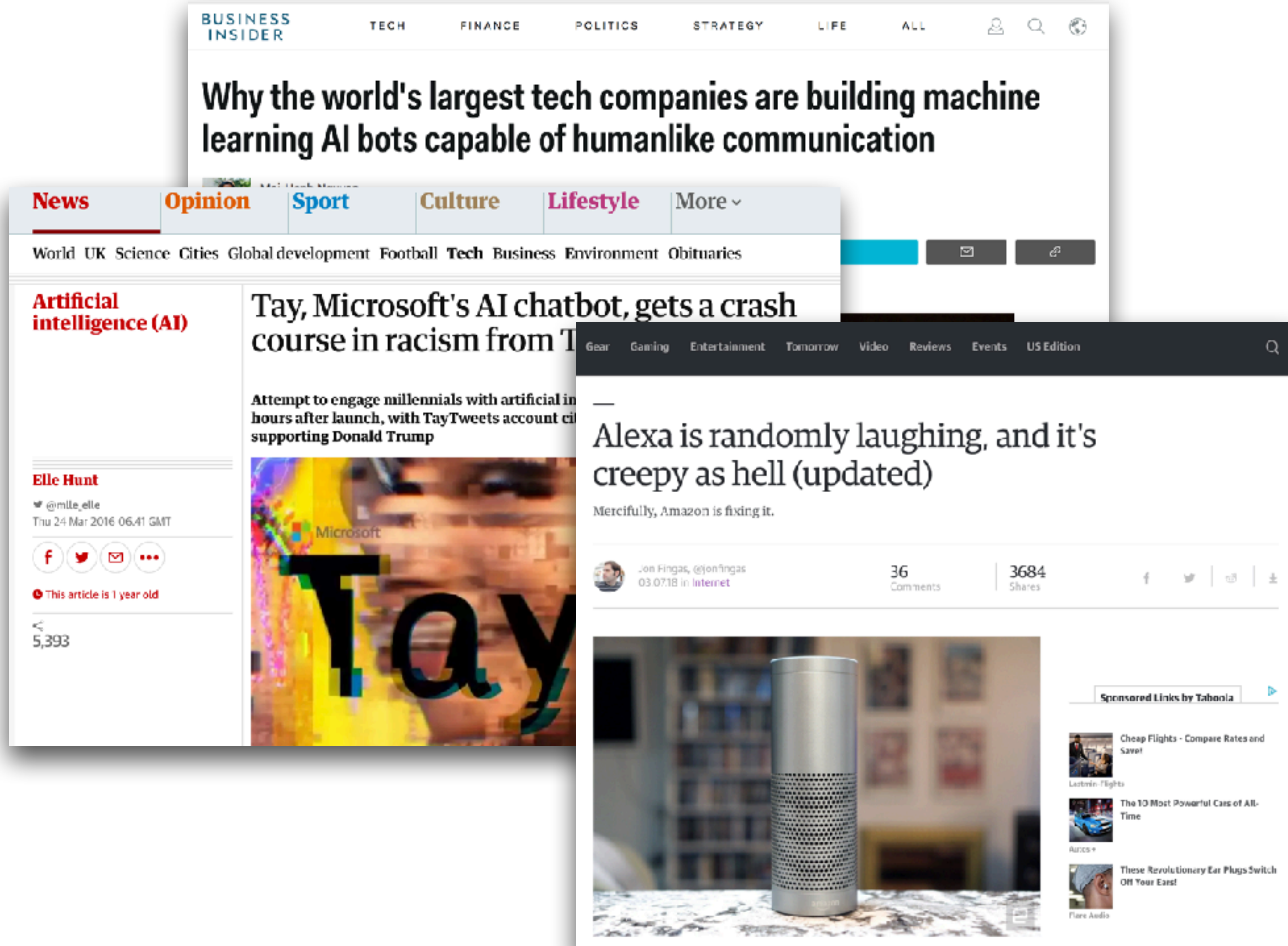
Conversational IR

An information retrieval system
with conversational interface
(in written or spoken form)



Similar to communication with a librarian:

- Understands natural language
- Elicits your information needs
- Knows your preferences


BUSINESS INSIDER

TECH FINANCE POLITICS STRATEGY LIFE ALL

Why the world's largest tech companies are building machine learning AI bots capable of humanlike communication

News Opinion Sport Culture Lifestyle More

World UK Science Cities Global development Football Tech Business Environment Obituaries

Artificial intelligence (AI)

Tay, Microsoft's AI chatbot, gets a crash course in racism from T

Attempt to engage millennials with artificial in
hours after launch, with TayTweets account ci
supporting Donald Trump

Elle Hunt

@mlle_elle

Thu 24 Mar 2016 06.41 GMT

This article is 1 year old

5,393

Gear Gaming Entertainment Tomorrow Video Reviews Events US Edition

Alexa is randomly laughing, and it's creepy as hell (updated)

Mercifully, Amazon is fixing it.

Jon Fingas, @jonfingas
03.07.18 in Internet

36 Comments

3684 Shares

Overview

1. Definitions

2. Methods

3. Evaluation

Overview

1 Definitions

- Task-oriented dialog agents
- Chatbots
- Conversational search

2 Methods

3 Evaluation

Task-oriented Dialog Agents

[Jurafsky and Martin 2017]

Hold short conversations, to get information from the user and help completing a task

- Usually explicitly model user intent and belief states
- Do not seek to sustain open-ended meaningful discourse
- E.g., Google now/home, Siri, Cortana, Alexa

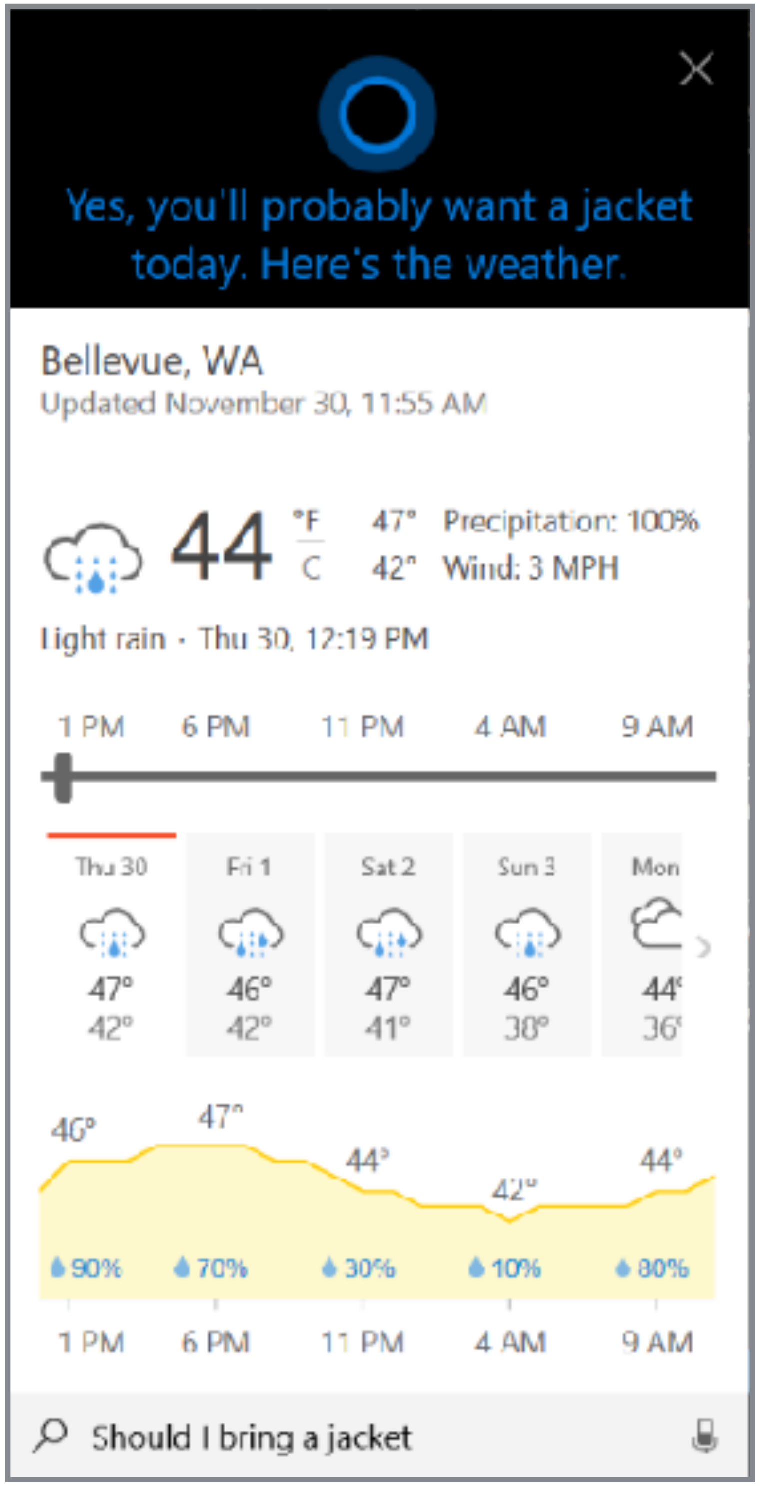
Example Dialog - Cortana

Human: Hey Cortana, what's the weather today?

Cortana: The forecast shows light rain with a high of 47 and a low of 42.

Human: Should I bring a jacket?

Cortana: Yes, you'll probably want a jacket today.

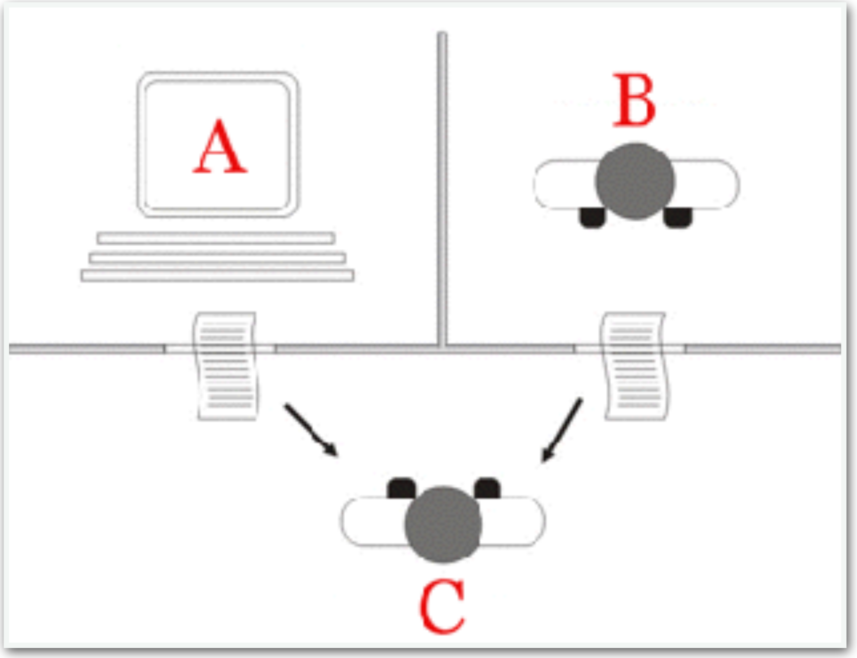


CHATBOTS

[Jurafsky and Martin 2017]

Hold extended conversations, to mimic the chat characteristics of human-human interactions.

- Often have entertainment value
- Attempt to pass Turing test
- E.g., Microsoft XiaoIce



Example Dialog - XiaoIce

Human: There is this guy I like a lot and I know he also has feelings for me. But he has ignored me for several days, what am I supposed to do?

Xiaoice: You just keep ignoring him, too.

Human: I just can't.

...

Xiaoice: Then you want a flash marriage? Well, I think he is not that into you.

Human: It's not a flash marriage for we have fallen into love for more than one year.

Xiaoice: It is a flash marriage.

Conversational Search

[Radlinski and Craswell 2017]

A system for **retrieving information** that permits a **mixed-initiative** back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both **short- and long-term knowledge** of the user.

RRIMS Properties

- User **R**evealment:
  System helps the user express/discover their true information need
- System **R**evealment:
  System reveals to the user its capabilities (what it can and cannot do)
- Mixed **I**nitiative:
  System and user both can take initiative as appropriate.
- **M**emory:
  User can reference past statements
- **S**et Retrieval:
  System can reason about the utility of sets of complementary items.

USER EXPECTATIONS

- A truly intelligent conversational search system does not exist.
- Users interaction and expectations are not known for us.
- Knowing users expectations is critical for the **design**, **evaluation**, and **improvement** of conversational search systems.

? Question

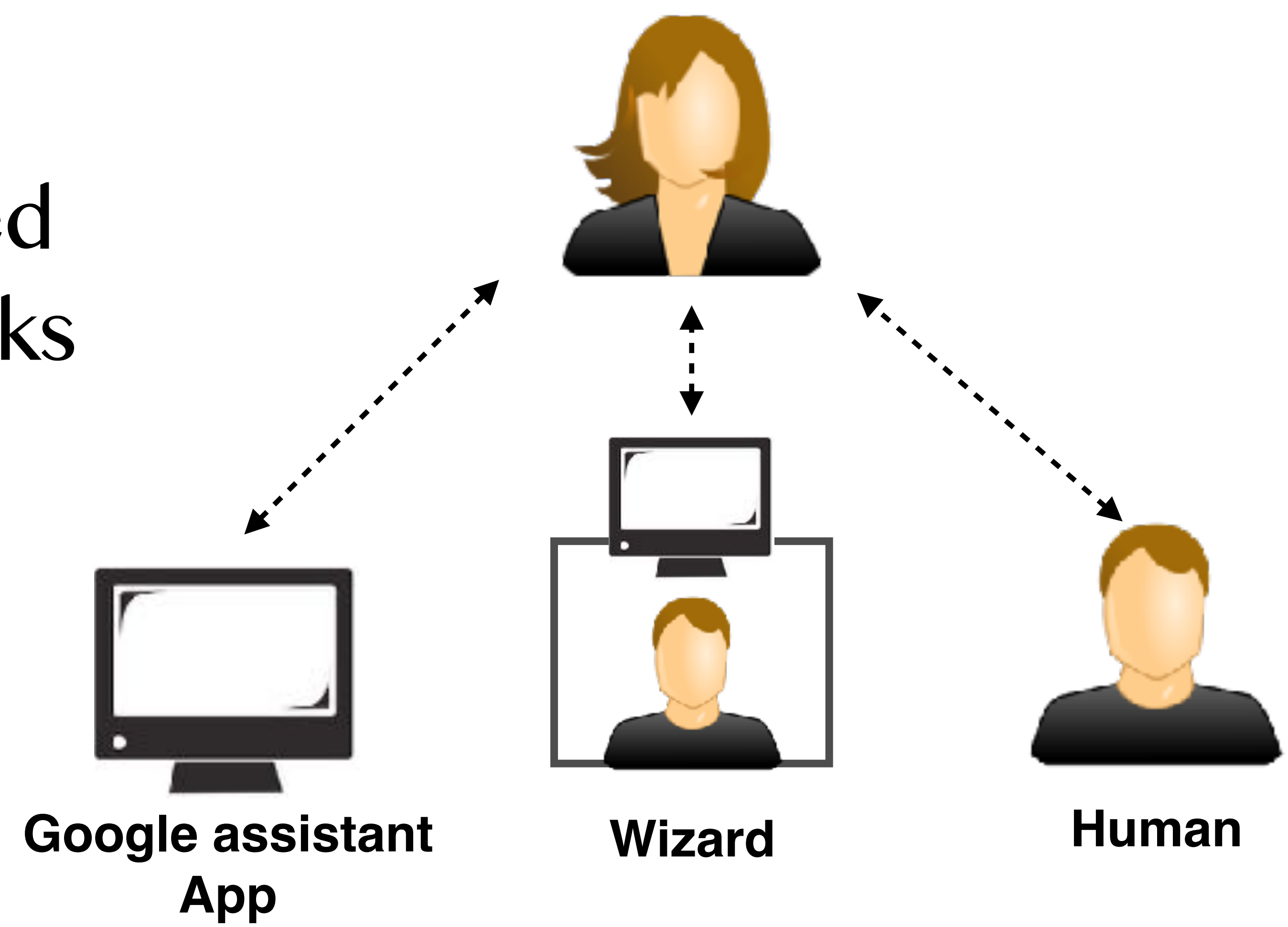
What are users expectations when interacting with a truly intelligent conversational search system?

USER EXPECTATIONS

[Vtyurina et al. 2017]

Experiments for identifying user expectations:

- 3 complex search tasks (from TREC Session track)
- 3 conversational agents
- A questionnaire was filled after completing the tasks



User Expectations

- **Maintaining context:**
  Enables search questions

- **Providing sources of answers:**
  Absence of trustworthy sources diminishes system credibility

- **Use of feedback:**
  Helps to back up from failure and improve results.

- **Opinion aggregation**
  Summary of different opinions is helpful for the users.

- **Direct answers vs. expanded information**
  User preferences vary on this matter

Overview

1. Definitions
2. Methods
   - Task-oriented dialog agents
   - Chatbots
3. Evaluation

Frame-based Architecture

- Based on the architecture of GUS system [Bobrow et al. 1977]
- **Domain ontology** represents the kinds of intentions the system can extract from user sentences
- A set of slots (**frame**), specifies what the system needs to know
- Each **slot** is filled with a value of a particular semantic type

Slot	Type	Question
ORIGIN CITY	City	"From what city are you leaving?"
DESTINATION CITY	City	"Where are you going?"
DEPARTURE TIME	Time	"When would like to leave?"
ARRIVAL TIME	Time	"When do you want to arrive?"

Frame-based Architecture

- A control structure is designed around the frame
- Often a Finite State Automata (FSA) is used

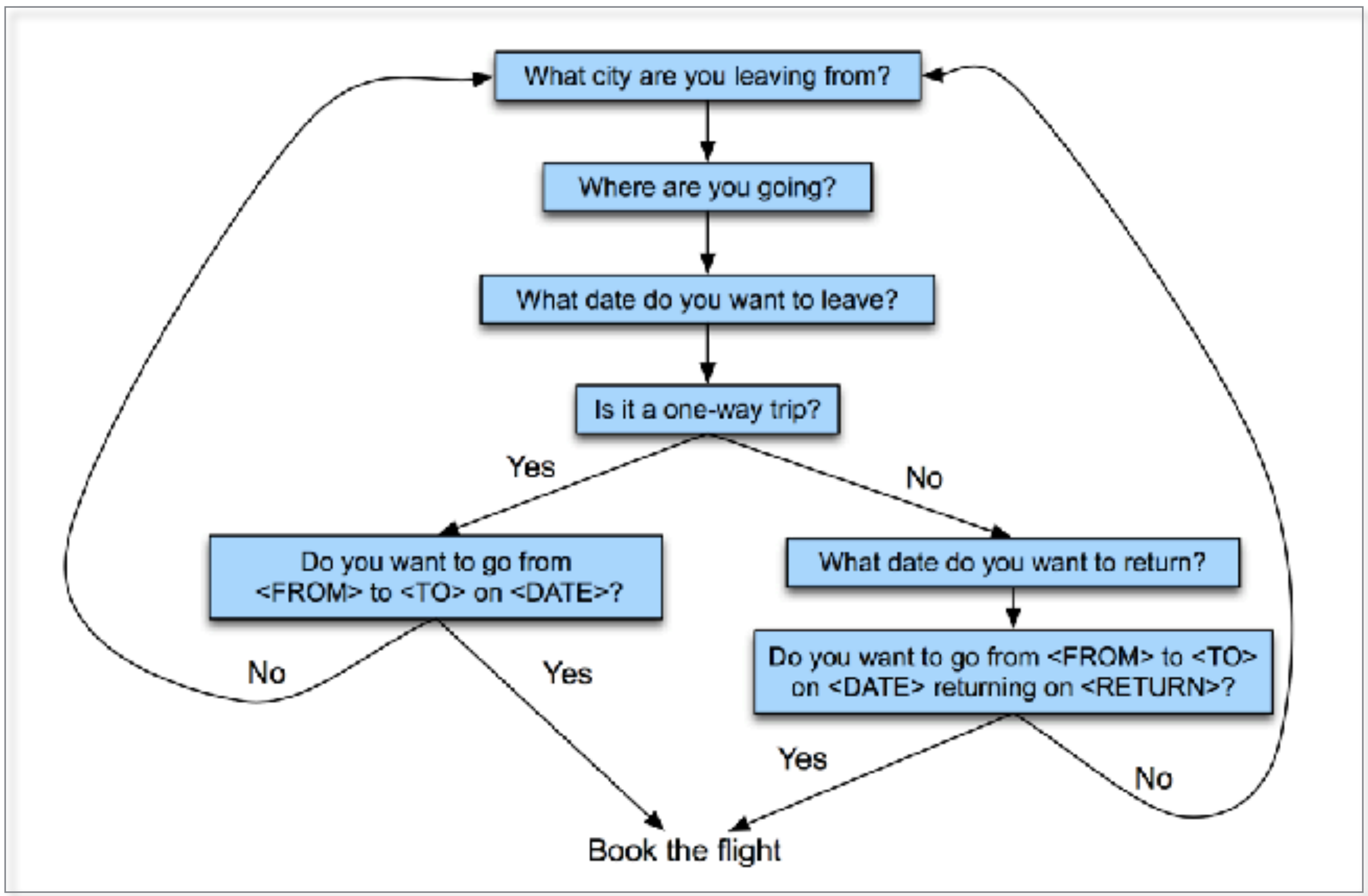


image: [Jurafsky and Martin 2017]

Frame-based Architecture

The frame structure of modern systems is flexible:

- Support multiple domain (e.g., hotel booking, route information)
- Allow mixed-initiative (not only system-initiative)
- Allow users switching between the frames
- Slots may be filled out of sequence
  - Multiple slots or nothing may be filled by an answer
  - Skips questions associated with slots that are already filled

FILING THE SLOTS

1) Domain classification

- Which Domain the user is talking about?
- E.g., dealing with calendar, booking a trip, or buying a house

2) Intent determination

- Which task the user is trying to accomplish?
- E.g., removing a calendar event, or show a flight

3) Slot filling

- Extracting slots and fillers from users' utterances

Filing the Slots

Example:

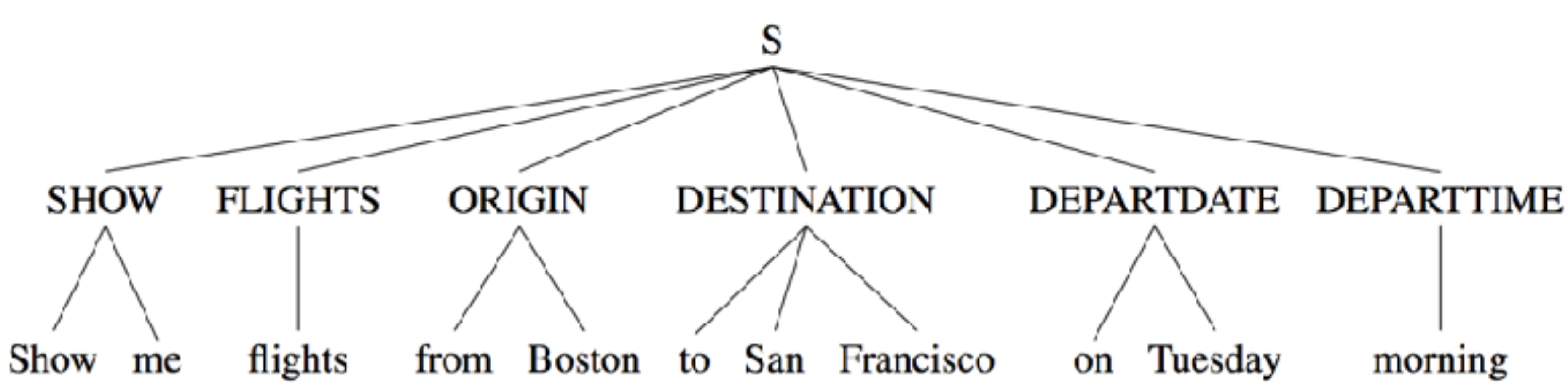
"Show me morning flights from Boston to
San Francisco on Tuesday"

DOMAIN:	AIR-TRAVEL
INTENT:	SHOW-FLIGHTS
ORIGIN-CITY:	Boston
ORIGIN-DATE:	Tuesday
ORIGIN-TIME:	morning
DEST-CITY:	San Francisco

FILING THE SLOTS

Rule based parsing:

- Hand written rules, even implemented with full grammars
- Context Free Grammar (CFG) parsing algorithms are often used
- Pros and cons:
  + High precision and sufficient coverage for narrow domains
  - Expensive and slow to create, low recall



Filing the Slots

Supervised machine learning:

- Train IOB tagger using a sequence model (e.g., CRF)
- Features:
  - Word embeddings
  - Word unigram, bigrams
  - Lexicon and slot transition features (e.g., DES after ORIGIN)

Example:

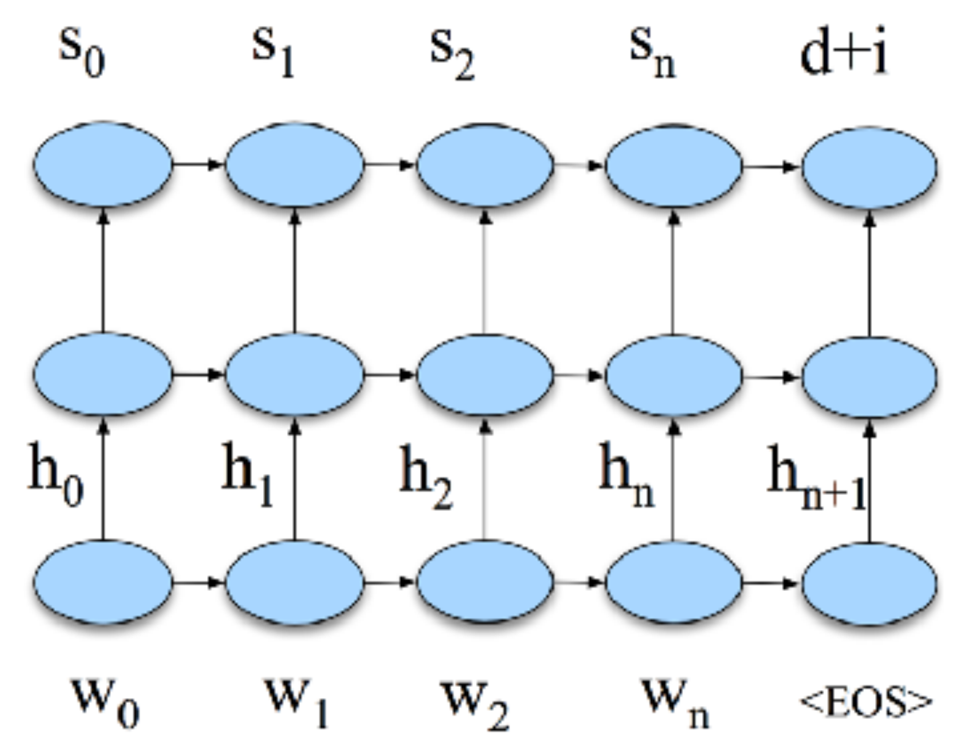
o	o	o	o	o	B-DES	I-DES	o	B-DEPTIME	I-DEPTIME	o
I	want	to	fly	to	San	Francisco	on	Monday	afternoon	please

Filing the Slots

Supervised machine learning:

- Train LSTM-style neural networks
- **Input:** a series of words (represented as embedding or 1-hot vectors)
- **Output:** IOB tags, plus the domain and intent

Supervised methods are often bootstrapped from rule-based systems



Overview

1. Definitions
2. Methods
   - Task-oriented dialog agents
   - Chatbots
     - Rule-based systems
     - Information retrieval techniques
     - Sequence to sequence techniques
3. Evaluation

Rule-based systems- Eliza

[Weizenbaum 1966]

- The most important chatbot in the history of the field
- Simulate a Rogerian psychologist
- People were deeply emotionally involved with Eliza
- Privacy concerns raised when people were told that ELIZA conversations may last for later analysis
- Some modern chatbots are based on updated versions of ELIZA's pattern/action architecture

Eliza- Example

Human: Men are all alike.

ELIZA: In what way?

Human: They're always bugging us about something or other.

ELIZA: Can you think of a specific example?

Human: Well my boyfriend made me come here.

ELIZA: Your boyfriend made you come here.

Human: He says I'm depressed much of the time.

ELIZA: I am sorry to hear you are depressed

..

Human: My father is afraid of everybody.

ELIZA: What else comes to mind when you think of your father?

Human: Bullies.

ELIZA: Does that have anything to do with the fact that your boyfriend made you come here?

ELIZA Algorithm

Memory trick

Non-committal response

Simplified algorithm of ELIZA:

function ELIZA GENERATOR(user *sentence*) **returns** *response*

Find the word *w* in *sentence* that has the highest keyword rank
  if *w* exists
    Choose the highest ranked rule *r* for *w* that matches *sentence*
    response ← Apply the transform in *r* to *sentence*
    if *w* = 'my'
      future ← Apply a transformation from the 'memory' rule list to *sentence*
      Push *future* onto memory stack
  else (no keyword applies)
   either
    response ← Apply the transform for the NONE keyword to *sentence*
   or
    response ← Pop the top response from the memory stack
 return(*response*)

Information Retrieval Techniques

- Mine human-human or even human-machine conversations
- Corpora:
  - Microblogging platforms; e.g., Twitter
  - Movie dialog corpora
  - Human responses to the chatbot (once the chatbot is used)
- Commercial implementation of the IR-based techniques is XioaIce

Information Retrieval Techniques

1) Return response to the most similar turn

$$r = response\left(\underset{t \in C}{\operatorname{argmax}} \frac{q^T t}{||q||t||}\right)$$

Turn: a single contribution to the dialog
consist of a single (or multiple) sentence(s)

- Finds turn t in corpus C that is most similar to query q
- Return the response to that turn

INFORMATION RETRIEVAL TECHNIQUES

2) Return the most similar turn

$$r = \underset{t \in C}{\operatorname{argmax}} \frac{q^T t}{||q|| \, ||t||}$$

While approach 1 is more intuitive, approach 2 (returning the most similar turn) seems to work better

SEQUENCE TO SEQUENCE TECHNIQUES

- Transducing from the user's prior turn to the system's turn
- Optimized to generate single responses
- Contentious coherence responses can be addressed using Reinforcement learning

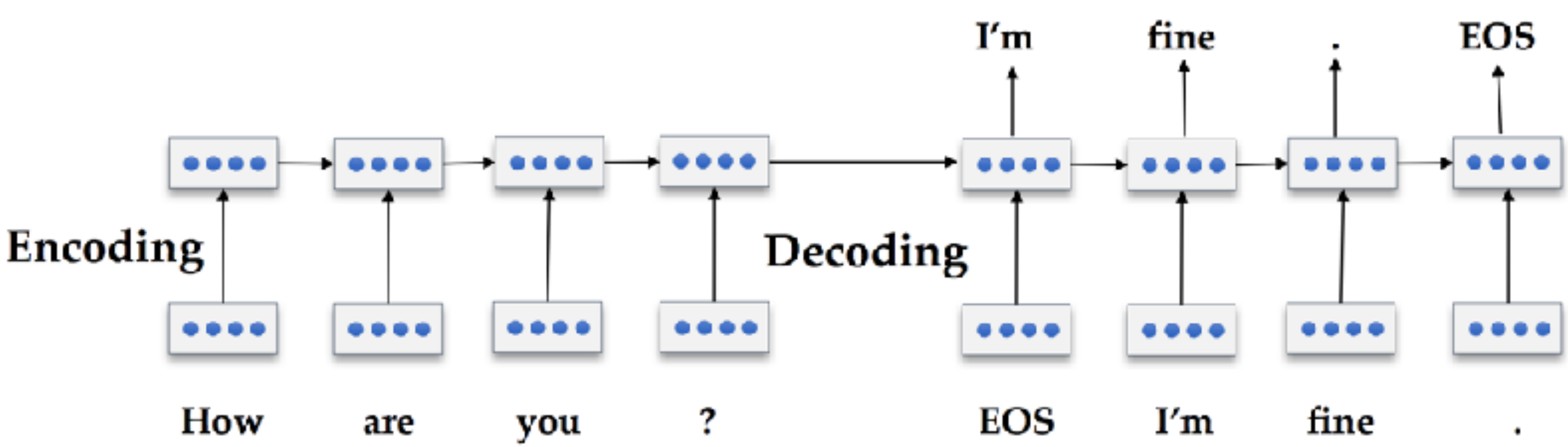


image: [Jurafsky and Martin 2017]

Overview

1. Definitions
2. Methods
3. Evaluation

Slot Filling Evaluation

1) Slot Error Rate for a sentence:

$$\text{Slot Error Rate} = \frac{\text{\# of inserted/deleted/subsituted slots}}{\text{\# of total reference slots for sentence}}$$

2) Task Error Rate:

- How often the task is done properly at the end of interaction
- E.g., Times that a correct meeting added to the calendar

Chatbot Evaluation

BLEU:

Measures word overlaps based on co-occurrences of n-grams in the ground truth and system responses.

Reference: Government officials are responsible for commuters

System: Responsible for commuters government formals

Metric	Value
Precision 1-gram	4/5
Precision 2-gram	2/4
Precision 3-gram	1/3
Brevity penalty (b)	5/6
BLEU-3	42 %

Geometric mean

$$\text{BLEU–n} = b.\left(\prod_{i=1}^{n} precision_i\right)^{\frac{1}{n}}$$

Chatbot Evaluation

Embedding Average:

1) Takes mean of the word embeddings of each token in a sentence:

$$\bar{e}_r = \frac{\sum_{w \in r} e_w}{|\sum_{w' \in r} e_{w'}|}$$

2) Compute the cosine similarity between their respective sentence level embeddings

$$\mathrm{EA} := \cos(\bar{e}_r, \bar{e}_{\hat{r}})$$

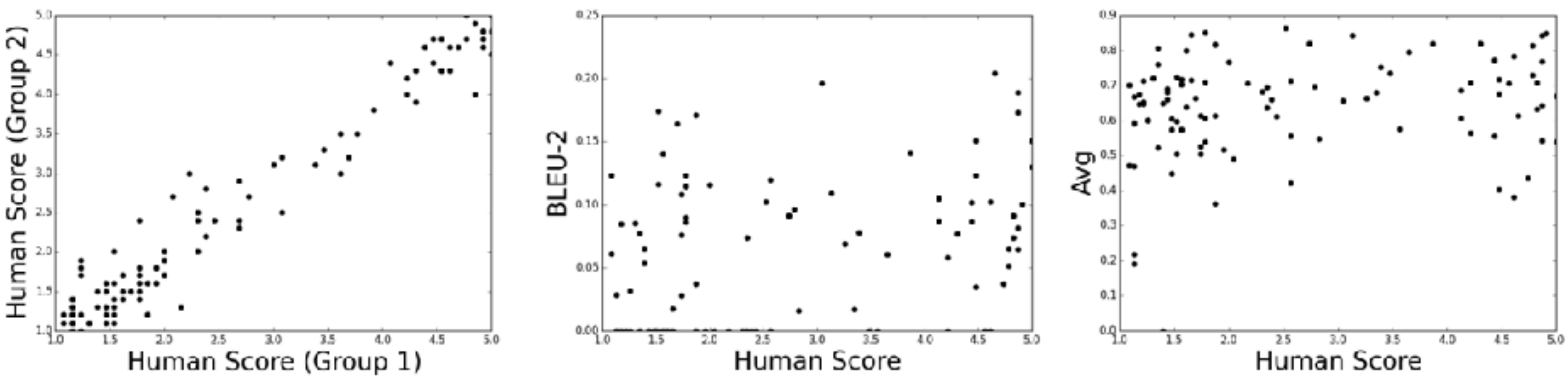
Human Evaluation

- Usually conducted using crowdsourced annotators
- Humans are asked to rate different aspects separately
  - E.g., 'adequacy', 'fluency' and 'informativeness' of the text
- The questions are of two types:
  - Compare the quality of system output responses pairwise e.g., "Decide which response is more informative."
  - Judge the response quality on a scale of (e.g., 1 to 5)

Chatbot Evaluation

[Liu et al. 2016]

Human evaluation does not correlate with automatic evaluation measures.

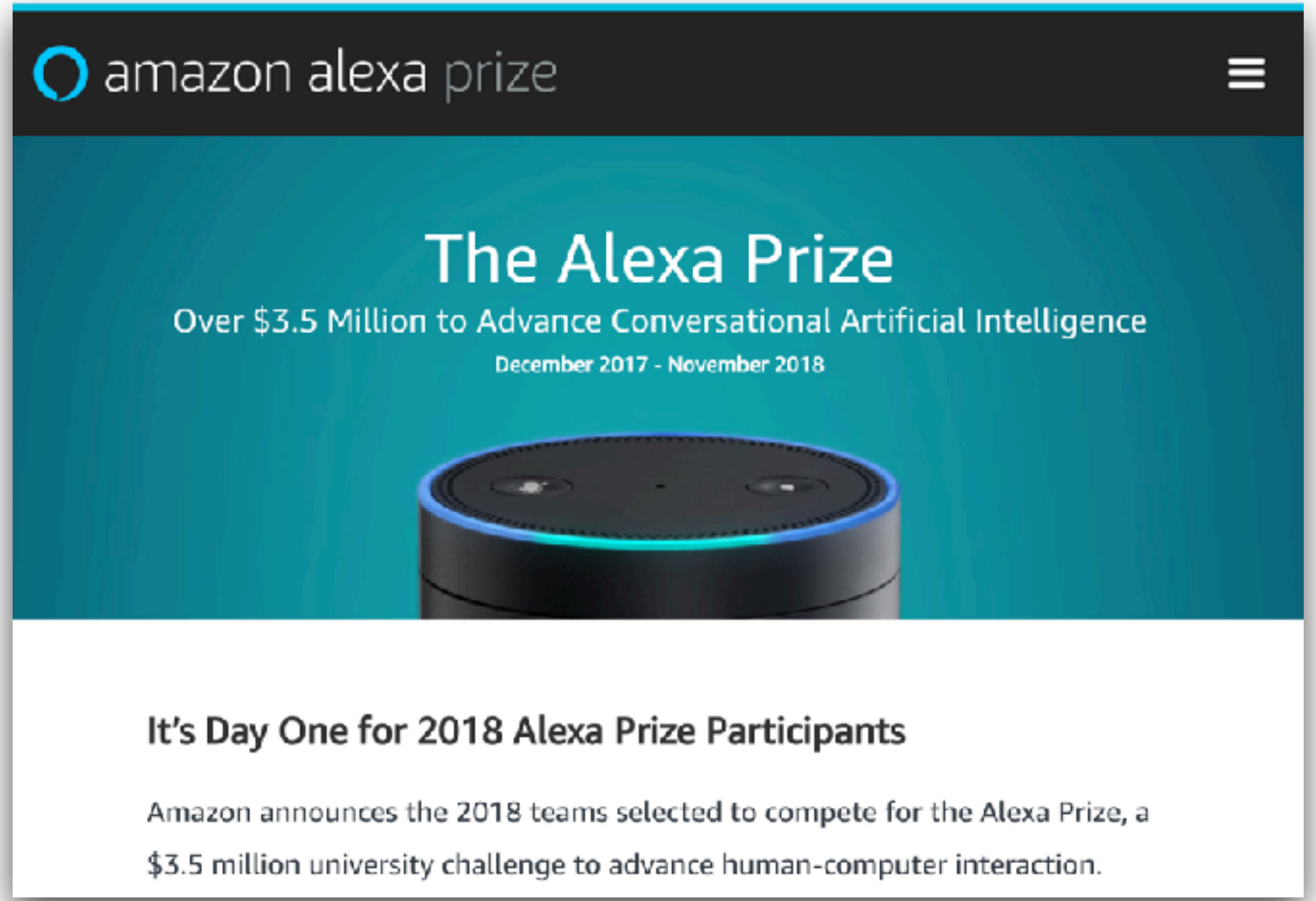


Conversational IR

Evaluation of conversational IR systems is an open question:

- Human evaluations are not reproducible
- Automatic evaluations are not representative
- Re-usable test collections are not available

The Alexa Prize



Wrapping up

- Definitions of task-based dialog systems, chatbots, and conversation IR systems
- System properties and user expectations of conversational search systems
- The frame-based architecture (used by most commercial dialog systems), and rule-/corpus-based chatbots
- Human evaluation vs. automatic evaluation

Future Directions

- Conversational IR is an exiting area to work on and it is currently in its infancy
- Can we make re-usable test collection with appropriate evaluation measures?
- Can a system mimic intelligent behavior of humans without having world knowledge?

Key References

- Jurafsky, D. and Martin, J. Speech and Language Processing. Prentice Hall, 2017.
- Eric, M. and Manning, C.D., (2017). A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue, EACL (2) 2017: 468-473
- Radlinski, F., & Craswell, N. (2017). A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (pp. 117–126).
- Vtyurina, A., Savenkov, D., Agichtein, E., & Clarke, C. L. A. (2017). Exploring Conversational Search With Humans, Assistants, and Wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2187–2193).
- Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., and Winograd, T. (1977). GUS, A frame driven dialog system. *Artificial Intelligence*, *8*, 155–173
- Weizenbaum, J. (1966). ELIZA – A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.
- K. Papineni, S. Roukos, T. Ward, and W. Zhu. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on Association for Computational Linguistics (ACL), 311-318.*
- Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Gao, J., Dolan, B., and Nie, J.-Y. (2015). A neural network approach to context-sensitive generation of conversational responses. In Proceedings of HLT-NAACL 2015: 196-205.
- Ghazvininejad, M., Brockett, C., Chang, M. W., Dolan, B., Gao, J., Yih, W. T., & Galley, M. (2017). A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932*.

THANK YOU