ACM UMAP 2020

28th Conference on User Modeling, Adaptation and Personalization

Predicting User Intents and Satisfaction with Dialogue-based Conversational Recommendations

Wanling CAI and Li CHEN

Department of Computer Science
Hong Kong Baptist University

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE
計算機科學系

Dialogue-based Conversational Recommender Systems (DCRS)

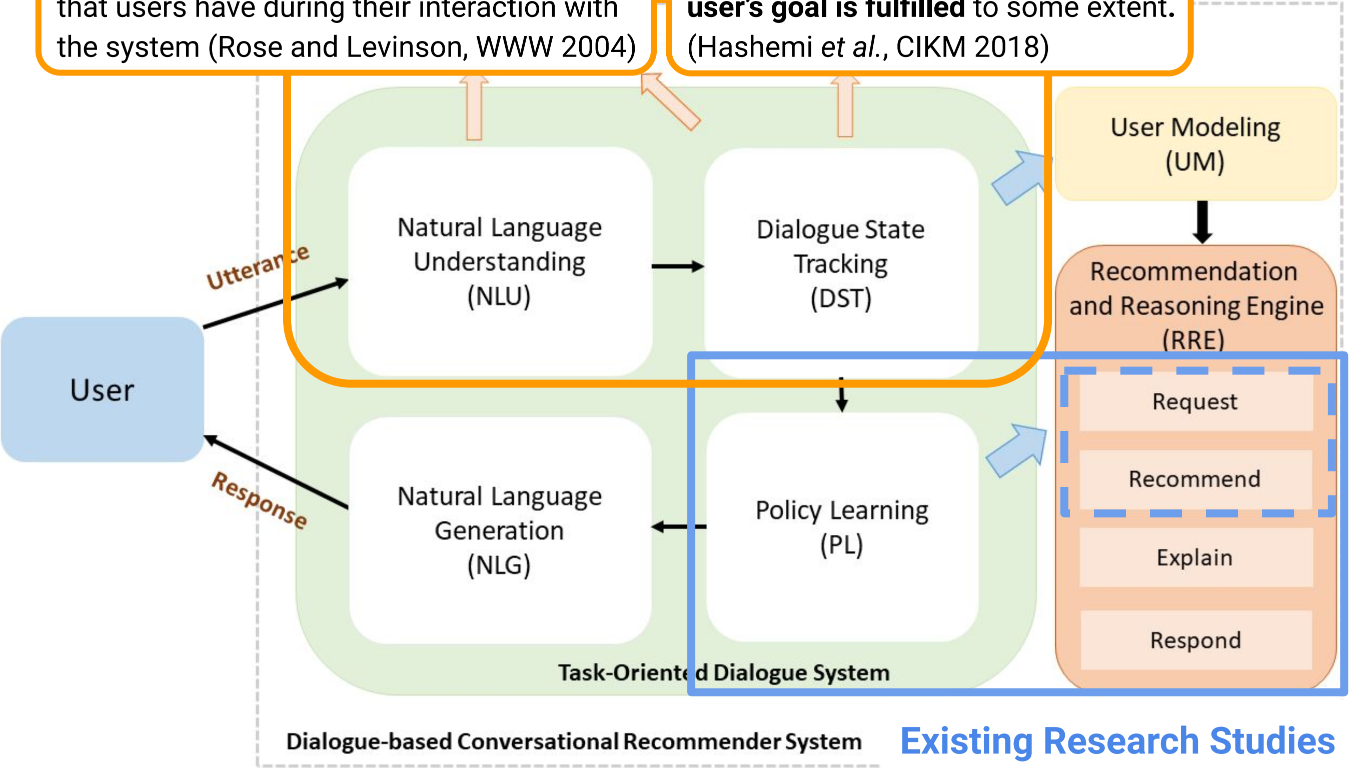
Dialogue-based Conversational Recommender System is one type of **task-oriented** dialogue system which **assists users in seeking for recommendations** (e.g., movies, music, hotels, and restaurants).

Icons made by: https://www.flaticon.com/

Dialogue-based Conversational Recommender Systems (DCRS)

User intent indicates the **goal** or **intention** that users have during their interaction with the system (Rose and Levinson, WWW 2004)

User satisfaction indicates **if the user's goal is fulfilled** to some extent. (Hashemi *et al.*, CIKM 2018)



Predicting user intents and satisfaction

Essential for DCRS

1. Understand users' preference
2. Select an appropriate system action
3. Adapt recommendation to user needs

Existing Research Studies

- **mainly focus on one-shot recommendation(s)**

Dialogue-based Conversational Recommender Systems (DCRS)

User Intent Discovery

- **Main idea:** investigate user intents/goals
- **Related Work**

Most-frequent user intents:
- *Recommendation*
- *Comparison*
- *Ask opinion*
- *Q&A*

Three session-aware intents:
- *Add filter condition*
- *See-more*
- *Negation*

(Yan *et al.*, AAAI 2017)

Identified from questions posted in the community sites

User initial query goals:
- *Objective,*
- *Subjective*
- *Navigational*

Follow-up query intents:
- *Refine*
- *Reformulate*
- *Start over*

(Kang *et al.*, Recsys 2017)

Identified from queries prompted by pre-defined system questions

Main limitation: The user data were not collected through natural conversations.

1st Research Objective:
- To understand the **dialogue-based interaction of users** by analyzing their conversations with human recommenders in a **multi-turn dialogue**.

Dialogue-based Conversational Recommender Systems (DCRS)

User Intent Prediction

- Utterance classification problem
- Previous work on conversational search and general dialogue systems

Classification Models:

Conventional Machine Learning Methods

- SVM (Bhargava *et al.*, ICASSP 2013)
- LR (Sun *et al.*, NIPS-SLU 2015)
- HMM (Surendran and Levow, SLP 2006)
- AdaBoost (Qu *et al.*, CHIIR 2019)

★ ***Advantages*:** Able to identify important features for user intent prediction.

Deep Learning Based Methods

- CNN (Bhargava *et al.*, ICASSP 2013)
- RNN/LSTM (Liu *et al.*, EMNLP 2017)

★ ***Advantages*:** Learn high-level features from utterances to improve prediction accuracy.

Features:

- Content
- Discourse
- Sentiment
- **Context (new)**

But few work studied user intent prediction specific to DCRS

- Lack of a well established taxonomy
- Lack of annotated dialogue data

2nd Research Objective:

- To define **various categories of feature** to predict user intents specific to DCRS.
- To investigate user intent prediction task in DCRS **using conventional ML methods and DL methods**.

Dialogue-based Conversational Recommender Systems (DCRS)

User Satisfaction Prediction

- Sequential classification problem
- Previous work on community question answering (CQA) and Intelligent assistant (IA)

However, few work investigated user satisfaction prediction specific to DCRS

Classification Models:

Conventional Machine Learning Methods

- SVM, Random Forests (Liu *et al.*, SIGIR 2008)
- LR (Mehrotra *et al.*, WWW 2019),
- GBDT (Kiseleva *et al.*, SIGIR 2016)

★ ***Advantages*:** Easy to interpret the reason of improved prediction performances

Deep Learning Based Methods

- LSTM/Bi-LSTM (Hashemi *et al.*, CIKM 2018)
- Neural Tensor Network (Chen *et al.*, WWW 2017)

★ ***Advantages*:** Better capture relationships within interaction sequences.

Features:

- Utterance-level features (i.e., content, discourse, sentiment features)
- Dialogue behavior features (i.e., user intents and recommender actions)

3rd Research Objective:

- To investigate the feasibility of leveraging dialogue behavior features (involving user intents and recommender actions) to predict user satisfaction with recommendations in DCRS.

Our Research Questions

RQ1: *How can we classify users' intents and recommenders' actions respectively in the dialogue conversation?*

RQ2: *How can we accurately predict a user's intents given her/his utterance in the recommendation dialogue?*

RQ3: *How does user satisfaction relate to their intents and recommender's actions in multi-turn interactions, and how can we accurately predict user satisfaction with the recommendation?*

Step 1: Taxonomy of User Intents & Recommender Actions

Recommendation Dialogue Data

ReDial Dataset

human-human dialogues centered around movie recommendations (Li *et al.*, NIPS 2018)

Seeker:	...
Recommender:	Another good one is Spaceballs.
Seeker:	**I did see that one, but I didn't really like it. I do love 80s movies though.**
Recommender:	Ok Well how about Planes, Trains and Automobiles.
Seeker:	**I may have seen that a long time ago but I can't remember. who is in that again?**
Recommender:	Steve Martin and John Candy. It is very funny.
Seeker:	**I love them both. I will try that one. Thanks so much!**

Statistics of our selected dialogue data (from ReDial)

Items	**SAT-Dial** (with user-satisfied recommendation)	**unSAT-Dial** (without user-satisfied recommendation)
# Conversations	253	83
# Human seekers	125 (# utterances: 1,711)	59 (# utterances: 550)
# Human recommenders	151 (# utterances: 1,747)	68 (# utterances: 575)
# Suggested movies per dialogue	4.57	4.51
# Turns per dialogue	mean=6.58, min=3, max=19	mean=6.49, min=3, max=12
# Words per utterance	mean=11.29, min=1, max=72	mean=10.72, min=1, max=69

ReDial dataset: https://redialdata.github.io/website/

Taxonomy of User Intents (RQ1)

Intent (Code)	Description	Percentage
Ask for Recommendation		**18.26%**
Initial Query (IQU)	Seeker asks for a recommendation in the first query.	12.91%
Continue (CON)	Seeker asks for more recommendations in the subsequent query.	3.10 %
Reformulate (REF)	Seeker restates her/his query with or without clarification/further constraints.	1.50%
Start Over (STO)	Seeker starts a new query to ask for recommendations.	0.84%
Add Details		**18.58%**
Provide Preference (PRO)	Seeker provides specific preference for the item s/he is looking for.	12.30%
Answer (ANS)	Seeker answers the question issued by the recommender.	4.91%
Ask Opinion (ASK)	Seeker asks the recommender's personal opinions.	2.39%
Give Feedback		**61.92%**
Seen (SEE)	Seeker has seen the recommended item before.	21.14%
Accept (ACC)	Seeker likes the recommended item.	18.89%
Reject (REJ)	Seeker dislikes the recommended item.	11.50%
Inquire (INQ)	Seeker wants to know more about the recommended item.	6.55%
Critique-Feature (CRI-F)	Seeker makes critiques on specific features of the current recommendation.	6.50%
Critique-Add (CRI-A)	Seeker adds further constraints on top of the current recommendation.	5.35%
Neutral Response (NRE)	Seeker does not indicate her/his preference for the current recommendation.	4.29%
Critique-Compare (CRI-C)	Seeker requests sth similar to the current recommendation in order to compare.	1.55%
Others	Greetings, gratitude expression, or chit-chat utterances.	14.55%

Taxonomy of Recommender Actions (RQ1)

Action (Code)	Description	Percentage
Request		**13.87%**
Request Information (REQ)	Recommender requests for the seeker's preference or feedback.	12.58%
Clarify Question (CLA)	Recommender asks a clarifying question for more details.	1.29%
Respond		**23.77%**
Respond-Feedback (RES)	Recommender responds to any other feedback from the seeker.	15.89%
Answer (ANS)	Recommender answers the question asked by the seeker.	7.88%
Recommend		**54.52%**
Recommend-Show (REC-S)	Recommender provides recommendation by showing it directly.	32.08%
Recommend-Explore (REC-E)	Recommender provides recommendation by inquiring about the seeker's preference	23.99%
Explain		**37.38%**
Explain-Introduction (EXP-I)	Recommender explains recommendation with non-personalized introduction.	22.83%
Explain-Preference (EXP-P)	Recommender explains recommendation based on the seeker's past preference.	13.01%
Explain-Suggestion (EXP-S)	Recommender explains recommendation in a suggestive way.	2.37%
Others	Greetings, gratitude expression, or chit-chat utterances.	29.80%

Step 2: User Intent Prediction

User Intent Prediction

- **Multi-label Classification Problem**
  For each given user utterance, the goal is to predict a subset of user intent labels.
  E.g., *"I did see that one, but I didn't really like it. I do love 80s movies though."*
  -> two intents: ***Reject*** and ***Critique-Add***
- **Methods**
  - **Classification Models**
    - 8 Machine Learning Models: LR, SVM, Naive Bayes, XGBoost, MLP, etc.
    - 2 Deep Learning Models: CNN, Bi-LSTM.
  - **Transformation Strategies** (transform multi-label classification into single-label problem)
    (1) Binary Relevance; (2) Classifier Chain; (3) Label Powerset.
- **Features**

Category	*Features*
Content	TF-IDF, Name Entity, # Relevant Items
Discourse	POS, 5W1H Question, Question Mark, Exclamation Mark, Utterance Length
Sentiment	Thanks, Sentiment Score, Opinion Lexicon
Context	Absolute Position, Utterance Similarity, Previous user intents & recommendation actions

- **Evaluation Metrics**
  - Label-based Accuracy
  - Precision
  - Recall
  - F1-score

Results - User Intent Prediction (RQ2)

Comparison of Classification Models

Methods	Binary Relevance				Classification Chain				Label Powerset			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Logistic Regression	0.5796	0.7160	0.6148	0.6612	0.6111	0.6898	0.6322	0.6596	0.6198	0.6791	0.6053	0.6400
SVM	0.5597	0.6701	0.6047	0.6332	**0.6293**	0.7179	0.6340	0.6730	0.6048	0.6004	0.6123	0.6056
Naive Bayes	0.4438	0.5137	0.5705	0.5400	0.4567	0.5137	0.5793	0.5439	0.5365	0.5989	0.5542	0.5755
Decision Tree	0.5264	0.5187	0.6778	0.5871	0.5356	0.5513	0.6325	0.5887	0.4515	0.4706	0.4755	0.4729
Random Forest	0.5742	0.5962	**0.7029**	0.6449	0.5968	0.6372	**0.6817**	0.6583	0.4794	0.4748	0.5096	0.4913
XGBoost	**0.5970**	**0.8169**	0.6007	**0.6919**	0.6274	**0.7957**	0.6268	**0.7010**	**0.6199**	**0.6868**	**0.6109**	**0.6463**
MLP	0.4773	0.7922	0.4743	0.5928	0.5079	0.7780	0.5045	0.6115	0.6157	0.6837	0.6029	0.6407

Methods	Acc	Pre	Rec	F1
ML-KNN	0.3960	0.4347	0.4335	0.4340
CNN	0.5698	0.6773	0.5618	0.6141
BiLSTM	0.5720	0.6747	0.5794	0.6234

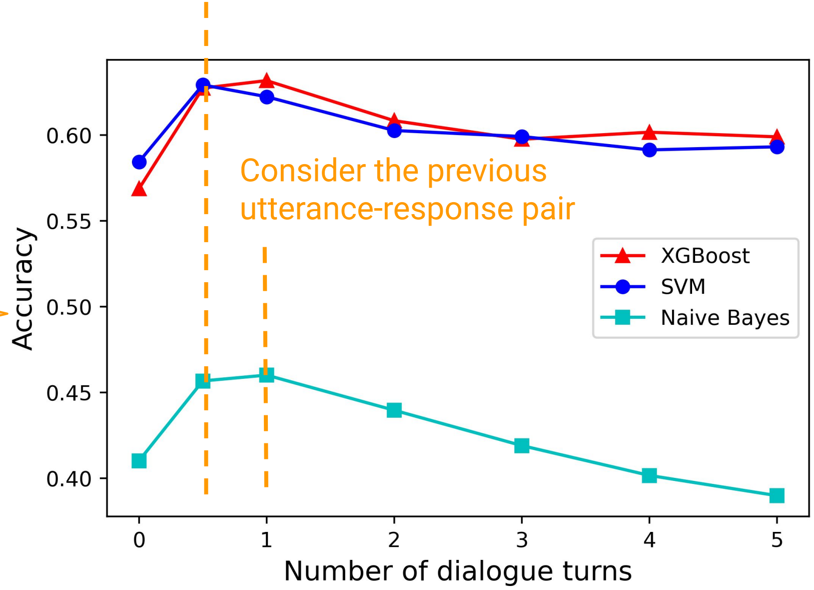
- ★ Classification Models: XGBoost (overall best)
- ★ Transformation Strategies: Classification Chain

Results - User Intent Prediction (RQ2)

Comparison of Feature Categories

	Cont	Disc	Sent	Context	Acc	Prec	Rec	F1
1 Category	✓				**0.4726**	**0.7165**	**0.4868**	**0.5793**
		✓			0.3918	0.5224	0.3841	0.4426
				✓	0.3407	0.5020	0.3343	0.4011
			✓		0.1993	0.3241	0.2044	0.2498
2 Categories	✓			✓	**0.5603**	**0.7669**	**0.5627**	**0.6488**
		✓		✓	0.5438	0.6946	0.5346	0.6039
	✓	✓			0.5291	0.7381	0.5350	0.6201
	✓		✓		0.4921	0.7289	0.5067	0.5972
			✓	✓	0.4587	0.6209	0.4518	0.5229
		✓	✓		0.4268	0.5553	0.4208	0.4787
3 Categories	✓	✓		✓	**0.6119**	**0.7913**	**0.6112**	**0.6896**
	✓		✓	✓	0.5870	0.7760	0.5887	0.6692
		✓	✓	✓	0.5698	0.7188	0.5569	0.6275
	✓	✓	✓		0.5415	0.7418	0.5500	0.6313
All	✓	✓	✓	✓	**0.6274**	**0.7957**	**0.6268**	**0.7010**

Only consider the previous recommender response



★ Content features → most effective

★ + Context features can significantly boost the prediction performance

★ Each feature category brings certain contribution

Results - User Intent Prediction (RQ2)

Individual Intent Prediction

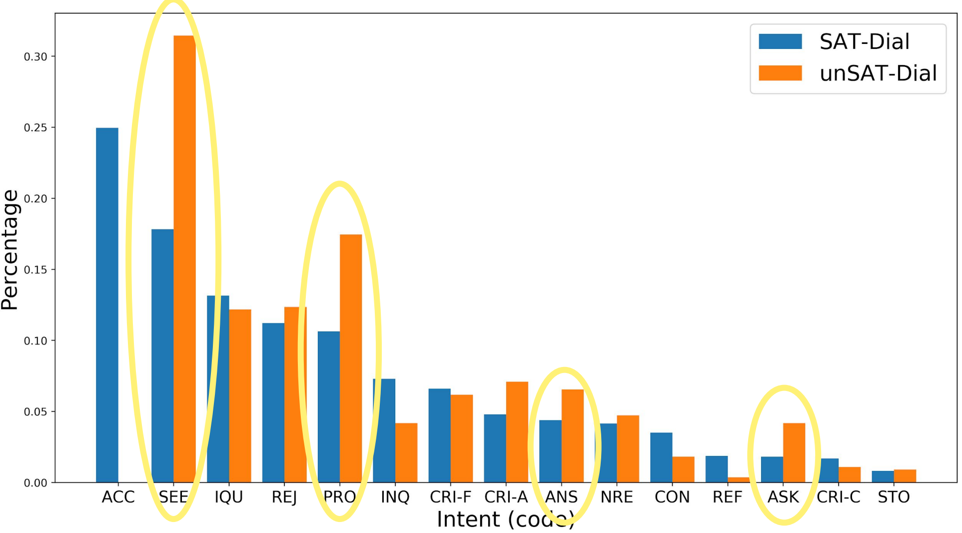
Intent Code	Cont	Disc	Sent	Context	Prec	Rec	F1
OTH	✓	✓	✓	✓	0.9325	0.9134	0.9224
IQU	✓	✓		✓	0.8985	0.8933	0.8941
SEE *Seen*	✓	✓	✓	✓	0.7859	0.6798	0.7270
ACC *Accept*	✓	✓	✓	✓	0.8391	0.6416	0.7239
CON	✓			✓	0.8014	0.5429	0.6294
INQ	✓	✓	✓	✓	0.6910	0.5352	0.5923
PRO *Provide Preference*	✓		✓	✓	0.7302	0.4930	0.5821
ANS	✓	✓		✓	0.6182	0.5053	0.5471
REJ *Reject*	✓	✓	✓		0.6704	0.4500	0.5357

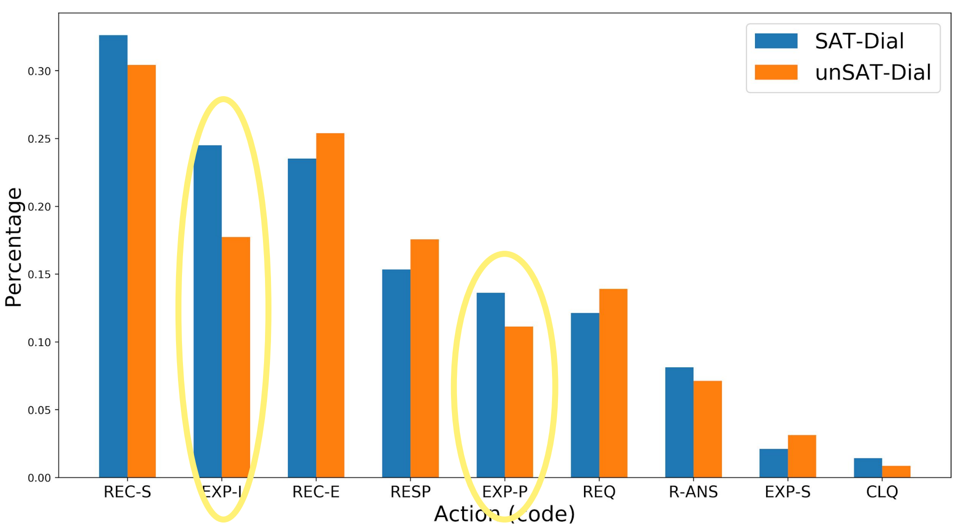
achieve relatively high accuracy

It is still challenging to identify some intents, e.g., *Provide Preference, Reject, Critiquing-Feature, Critiquing-Add.*

Step 3: User Satisfaction Prediction

Dialogue Data Analysis





Distribution comparison between satisfactory (SAT-Dial) and unsatisfactory dialogues (unSAT-Dial)

Seekers more often add details to indicate their preferences in unSAT-Dial

User Intents

- **unSAT-Dial:** *See, Add Details (i.e., Provide Preference, Answer, and Ask)*
- **SAT-Dial:** *Inquire*

Providing explanations is likely to increase users' acceptance

Recommender Actions

- **SAT-Dial:** *Explain (e.g., Explain-Introduction, Explain-Preference)*

User Satisfaction Prediction

- **Binary Classification Problem**
  Given a fixed number (N) of turns in the dialogue, the goal is to predict if the user would eventually accept a recommendation.

- **Classification Models**
  - 8 Machine Learning Models: LR, SVM, Naive Bayes, XGBoost, MLP, etc.

- **Features**
  - Dialogue behavior features (i.e., user intents and recommender actions)
  - Utterance-level features (i.e., content, discourse, and sentiment features)

- **Evaluation Metrics**
  - Precision
  - Recall
  - F1-score

Category	*Features*
Content	TF-IDF, Name Entity, # Relevant Items
Discourse	POS, 5W1H Question, Question Mark, Exclamation Mark, Utterance Length
Sentiment	Thanks, Sentiment Score, Opinion Lexicon

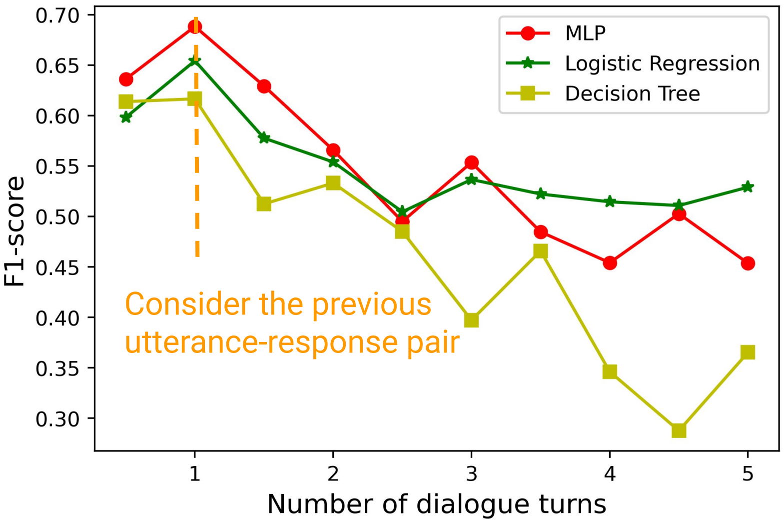
Results - User Satisfaction Prediction (RQ3)

Comparison of Classification Models

Methods	Cont	Disc	Sent	Dial	Prec	Rec	F1
Logistic Regression	✓	✓		✓	0.8488	**0.5806**	0.6795
SVM		✓		✓	0.8778	0.5556	0.6629
Naive Bayes				✓	0.8833	0.5556	0.6651
Decision Tree				✓	0.7109	0.5528	0.6167
Random Forest				✓	0.8862	0.5306	0.6503
XGBoost				✓	0.7897	0.5653	0.6426
MLP				✓	**0.8990**	0.5681	**0.6884**
KNN			✓	✓	0.8850	0.5181	0.6427

Comparison of Feature Categories

Method	Cont	Disc	Sent	Dial	Prec	Rec	F1
MLP				✓	**0.8990**	**0.5681**	**0.6884**
	✓				0.6551	0.4944	0.5501
		✓			0.5570	0.3486	0.4122
			✓		0.6067	0.2681	0.3606
	✓	✓	✓	✓	0.7995	0.5444	0.6292



- ★ Classification Models: MLP (best precision & F1)
- ★ Effective Features:
    - Dialogue behavior features (i.e., user intents and recommender actions)

Conclusions

Summary

1. *Two hierarchical* ***taxonomies established for user intents and recommender actions*** *respectively*
2. ***User intent prediction****: Some methods (such as XGBoost and SVM) can achieve outperforming accuracy by unifying* ***four feature categories (i.e., content, sentiment, discourse, and context)***
3. ***User satisfaction prediction****: Leveraging* ***both user intents and recommender actions*** *enables some model like MLP to achieve competitive accuracy*

Intent Annotation of Recommendation Dialogue (IARD) dataset *is publicly available:*

https://github.com/wanlingcai1997/umap_2020_IARD.git

Future Work

1. *To verify the taxonomies' generalizability to other dialogues and product domains*
2. *To label more dialogue data and identify whether deep learning (DL) methods would become superior when the dataset is enlarged*
3. *To investigate the temporal sequence of utterances/responses within a dialogue, which might act as potentially useful context features to further improve the prediction accuracy*

Thanks! Q&A

Wanling Cai
cswlcai@comp.hkbu.edu.hk

Li Chen
lichen@comp.hkbu.edu.hk

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系