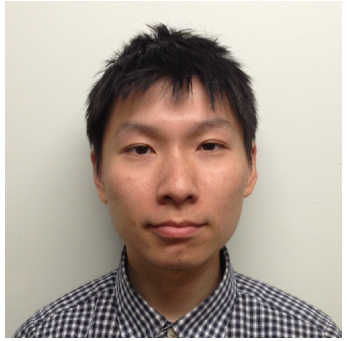
Talking to Your TV: Context-aware Voice Search with Hierarchical Recurrent Neural Networks

CIKM'17 Singapore Nov. 6th 2017

Jinfeng Rao
Univ. of Maryland

Ferhan Ture
Comcast Applied AI Lab

Hua He
Univ. of Maryland



Oliver Jojic
Comcast Applied AI Lab

Jimmy Lin
Univ. of Waterloo

Comcast's Voice Interface for TV

22 million customers in 40 states of United States since 2015
http://www.xfinity.com/voice-remote

1

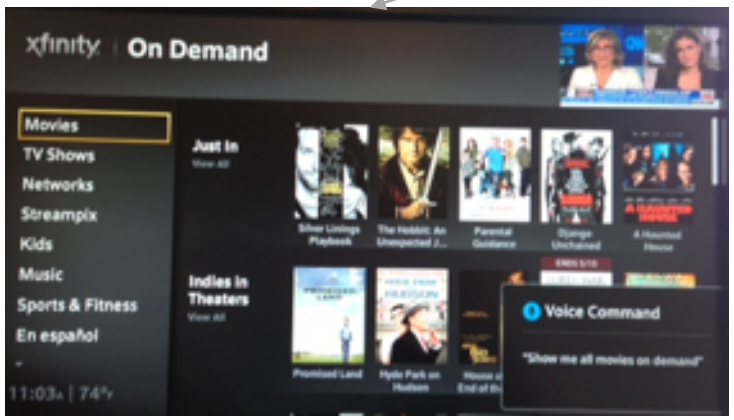
NLP IN THE CLOUD

2 SPEECH-TO-TEXT

3 INTENT UNDERSTANDING

4 ACTION RESOLUTION

5 MESSAGE MANAGEMENT

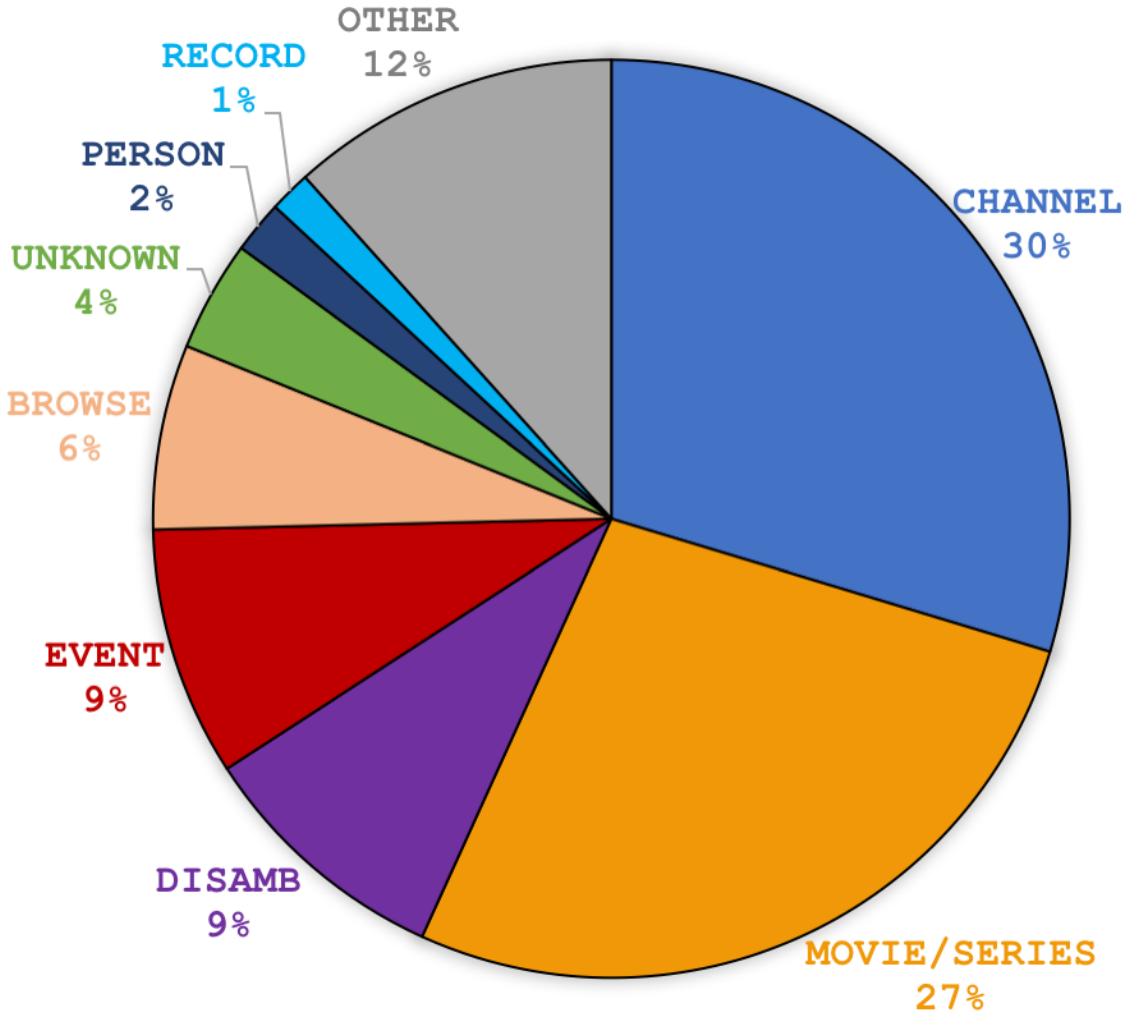


COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

What do user say to their TV?



1. Channel (30%): "HBO"
2. Movie/Series(27%): "Find game of throne"
3. Event (9%): "Oscar show"
4. Browse(6%): "free action movies"
5. Hundreds of intents more……

COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Challenges

- Short length of voice queries
  - Analysis over a week of 32M voice queries shows the average length is 2.04 words, much shorter than published stat. on smartphone & computers [1, 2].
- Ambiguity
  - Query-level: a query "Chicago fires" can refer to either a television series or a soccer team.
  - Program-level: many TV programs share similar names: *The Princess Diaries, The Princess Diaries II: Royal Engagement, The Princess Knight, etc…*
- Many speech recognition errors
  - Little query/program lexical overlap because of SR errors, such as query "you" to program "Calliou"

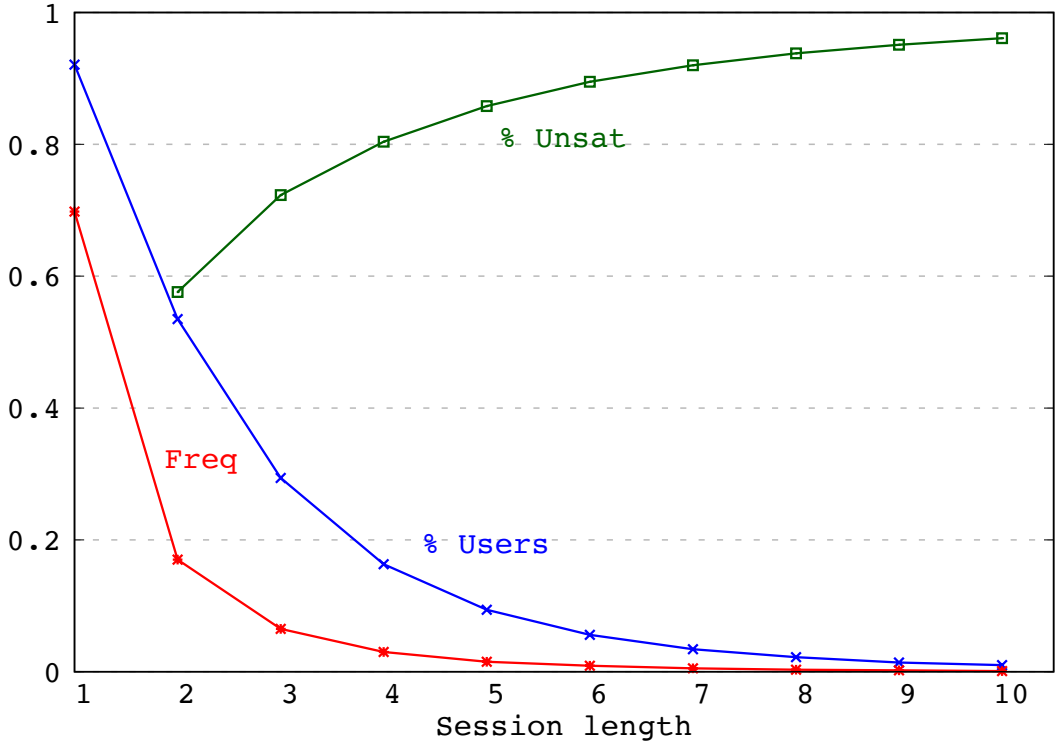
COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Session Analysis (~32M voice queries)

- For each session length, we plot three values: frequency of sessions (red), percentage of users that issued at least one session of that length (blue), percentage of unsatisfactory users (green).



- From the red curve, about 30% sessions have multiple queries, accounting 57% of total queries.
- From the blue curve, more than 50% users have issued at least one multiple-query session.
- From the green curve, the unsatisfactory rate keeps going up as the session length increases.
- Overall, session modeling is very important for improving user experience!

COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Problem Formulation

- Assumption
  - User will keep issuing queries until he find the intended program, i.e., ["hbo series", "game of throw", "game of throne"]

- Given a voice query session $S_i = [q_1, q_2, \ldots, q_n]$, we aim to predict the intended program at each query time, exploiting previous queries as contexts:

$$\text{Data: } D = \{(s_i, p_i) \mid s_i = [q_{i_1}, \ldots, q_{i_{|s_i|}}],\ p_i \in \Phi\}_1^{|D|}$$

$$\text{Model: } \hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{|D|} \prod_{t=1}^{|s_i|} P(p_i | q_{i_1}, \ldots, q_{i_t}; \theta)$$

where S_i denotes a voice session, P_i is the ground truth program,

COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Problem Decomposition

$$\text{Model: } \hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{|D|} \prod_{t=1}^{|s_i|} P(p_i | q_{i_1}, \ldots, q_{i_t}; \theta)$$

$$P(p_i | q_{i_1}, \ldots, q_{i_t}) = P(p_i | c_{i_t}) \cdot P(c_{i_t} | v_{i_1}, \ldots, v_{i_t}) \cdot \prod_{j=1}^{t} P(v_{i_j} | q_{i_j})$$

$$v_{i_t} \sim \mathbb{F}(q_{i_t}; \theta_{\mathbb{F}}), \quad c_{i_t} \sim \mathbb{G}(v_{i_1}, \ldots, v_{i_t}; \theta_{\mathbb{G}}), \quad p_i \sim \mathbb{H}(c_{i_t}; \theta_{\mathbb{H}})$$

$$1 \leq t \leq |s_i|$$

- F: embedding function that converts a query string to a embedding vector
- G: contextual function that converts the first t-th queries into a contextual embedding
- H: classification function that maps a contextual embedding to a program Pi

COMCAST

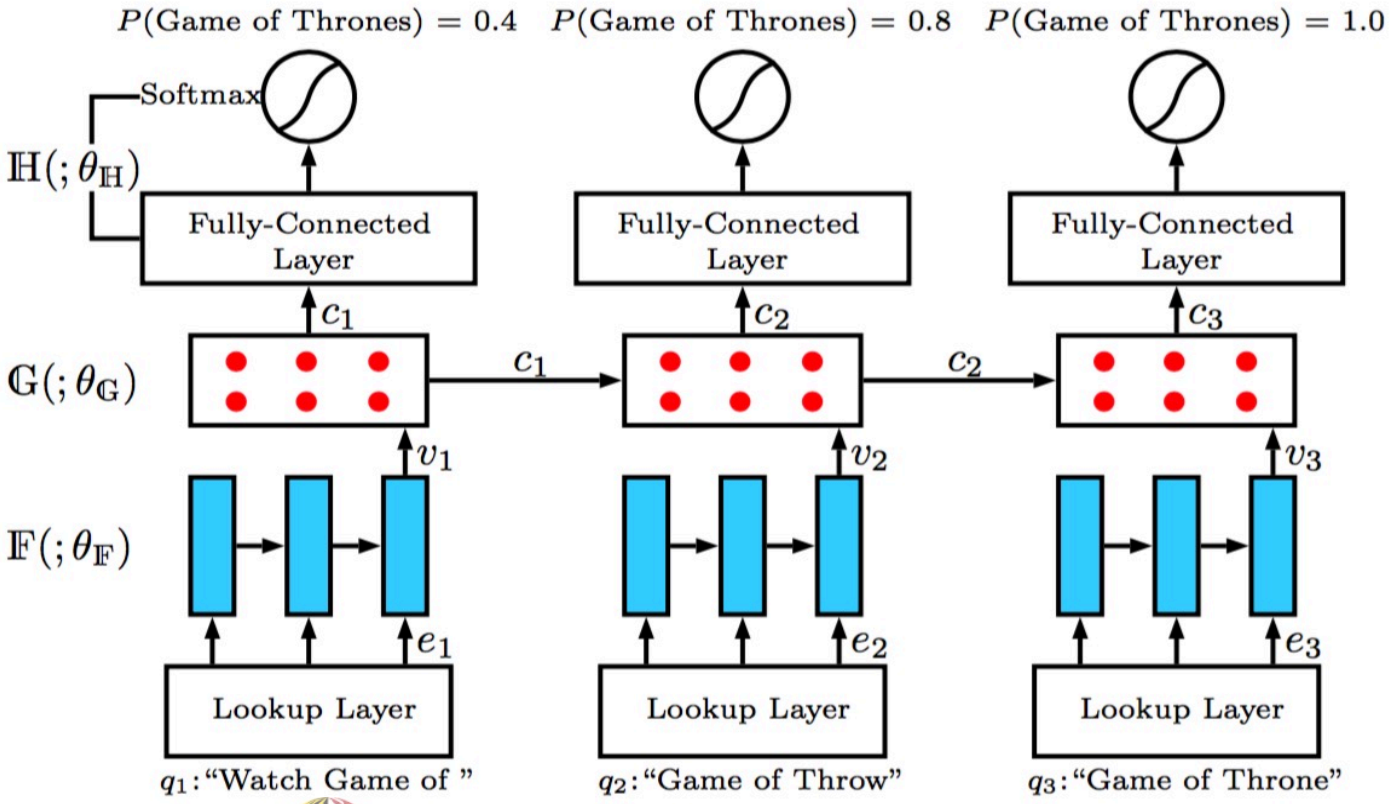
UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Model: Hierarchical Recurrent Neural Networks

$$v_{i_t} \sim \mathbb{F}(q_{i_t}; \theta_{\mathbb{F}}), \quad c_{i_t} \sim \mathbb{G}(v_{i_1}, \ldots, v_{i_t}; \theta_{\mathbb{G}}), \quad p_i \sim \mathbb{H}(c_{i_t}; \theta_{\mathbb{H}})$$
$$1 \leq t \leq |s_i|$$

- Two hierarchical LSTMs to model the embedding function F and contextual function G
- A multi-layer perceptron to model the classification function G.



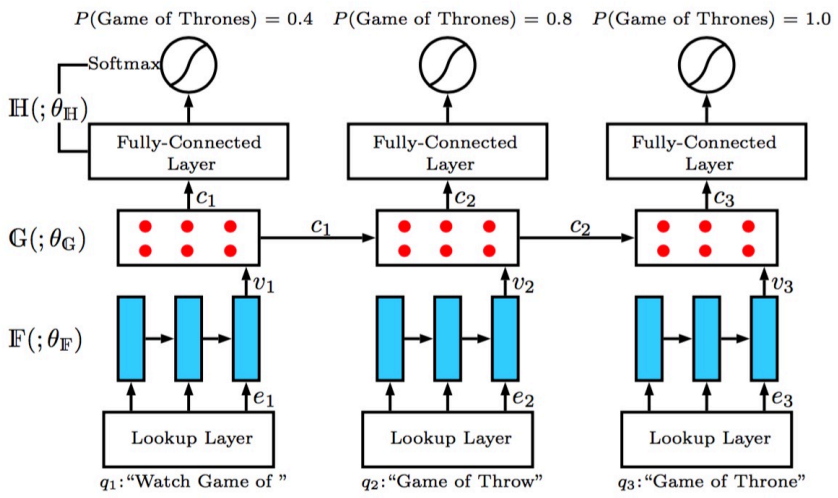
COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Model Variants

- Three ways of query representations
  - **character-level:** better for resolving speech recognition errors (i.e., "you" to "Caillou")
  - **word-level**: better for semantic modeling
  - **combined**: combine the above two embedding ways



- Three model options
  - **basic:** remove the contextual modeling part
  - **full context:** full architecture
  - **constrained context:** full architecture, pretrain the underlying layers, then fine-tune the top layers

- Total 3*3 model configurations

COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Data Generation

- Collected ~32M raw voice queries during Feb. 22-28th 2016 from 2.5M unique viewers in the Comcast's Xfinity X1 platform.
- Selected 45 seconds to sessionize queries, yielding ~20M sessions.
- The ground truth labels of sessions are collected by viewers' watching behaviors after their queries, yielding 13M session-program pairs.
- Limited query intents to be one of the {SERIES, MOVIE, MUSICVEDIO, SPORTS} categories, generating 2.1M labeled pairs

Dataset	#sessions	#queries	Avg. session len	Avg. query len
Train	126016	181058	1.44	2.34
Validation	28427	82828	2.91	2.30
Test	28173	82272	2.92	2.30

Training Details

- GloVe word embeddings
  - 1.8K/20.4K words were not found in the vocabulary
  - Unseen word embeddings were randomly initialized from [-0.05, 0.05]
- RMS-PROP algorithm for parameter updating
  - learning rate was set to 0.001 initially, then decreased a factor of 3 if validation set loss doesn't go down for three epoches.
- Categorical loss function was adopted
- At test time, we selected the model that obtained the highest P@1 accuracy at validation set for evaluation.

COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Metrics and Baselines

- Metrics
  - Precision at 1 (P@1), Precision at 5 (P@5)
  - Mean Reciprocal Rank (MRR)
  - Query Reduction (QR): number of queries saved by our methods
- Non-context Baselines
  - TF-IDF: match query with programs by character-level 3grams.
  - BM25
  - SVM-Rank with three sets of features: 1) BM25 score, 2) popularity prior of the program, and 3) the word embedding based similarity features.
  - Deep Structure Semantic Models (DSSM): 3gram based neural matching method
  - Basic model with three query representations.
- Context-aware Baselines
  - DSSM+S: concatenating queries in one session by whitespace.
  - Basic+S

COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Main Results

ID	Model	Query	P@1	P@5	MRR	QR
1	TF-IDF	3-gram	0.518^{10}	0.593^{10}	0.543^{10}	0.932^{10}
2	BM25	3-gram	$0.533^{1,10}$	0.596^{10}	$0.565^{1,10}$	$0.947^{1,10}$
3	SVM^{rank}	-	$0.547^{1,2,10}$	$0.621^{1,2,10}$	$0.582^{1,2,10}$	$0.962^{1,2,10}$
4	DSSM	3-gram	$0.568^{1,2,3,10}$	$0.617^{1,2,10}$	$0.584^{1,2,10}$	$1.001^{1,2,3,4,10}$
5	DSSM+S	3-gram	$0.550^{1,2,10}$	$0.618^{1,2,10}$	$0.576^{1,2,10}$	$0.972^{1,2,10}$
6	Basic	char	$0.605^{1\text{-}5,10}$	$0.647^{1\text{-}5,10}$	$0.690^{1\text{-}5,10}$	$1.108^{1\text{-}5,10}$
7	Basic	word	$0.609^{1\text{-}5,10}$	$0.644^{1\text{-}5,10}$	$0.677^{1\text{-}5,10}$	$1.086^{1\text{-}5,10}$
8	Basic	comb	$0.614^{1\text{-}5,9,10}$	$0.651^{1\text{-}5,10}$	$0.687^{1\text{-}5,10}$	$1.113^{1\text{-}5,9,10}$
9	Basic+S	comb	$0.596^{1\text{-}5,10}$	$0.645^{1\text{-}5,10}$	$0.697^{1\text{-}5,10}$	$1.061^{1\text{-}5,10}$
10	Context-f	char	0.482	0.532	0.580^{1}	0.856
11	Context-f	word	$0.599^{1\text{-}5,10}$	$0.638^{1\text{-}5,10}$	$0.687^{1\text{-}5,10}$	$1.075^{1\text{-}5,10}$
12	Context-f	comb	$0.598^{1\text{-}5,10}$	$0.643^{1\text{-}5,10}$	$0.688^{1\text{-}5,10}$	$1.039^{1\text{-}5,10}$
13	Context-c	char	$0.639^{1\text{-}12}$	$0.684^{1\text{-}12}$	$0.731^{1\text{-}12}$	$1.117^{1\text{-}7,9\text{-}12}$
14	Context-c	word	$0.639^{1\text{-}12}$	$0.683^{1\text{-}12}$	$0.729^{1\text{-}12}$	$1.112^{1\text{-}7,9\text{-}12}$
15	Context-c	comb	$\mathbf{0.643}^{1\text{-}12}$	$\mathbf{0.687}^{1\text{-}12}$	$\mathbf{0.734}^{1\text{-}12}$	$\mathbf{1.128}^{1\text{-}12}$

(b) multiple-query

Observations

- DSSM is the strongest baseline
- A simple query concatenation (DSSM+S, Basic+S) doesn't capture the session-aware context signal.
- Constrained context method (Context-c) performed the best, significantly better than all other methods.
- Full context method (Context-f) suffered, probably because of insufficient data for optimization at once.
- Char-level and word-level embedding perform closely, while combination-based is the best.

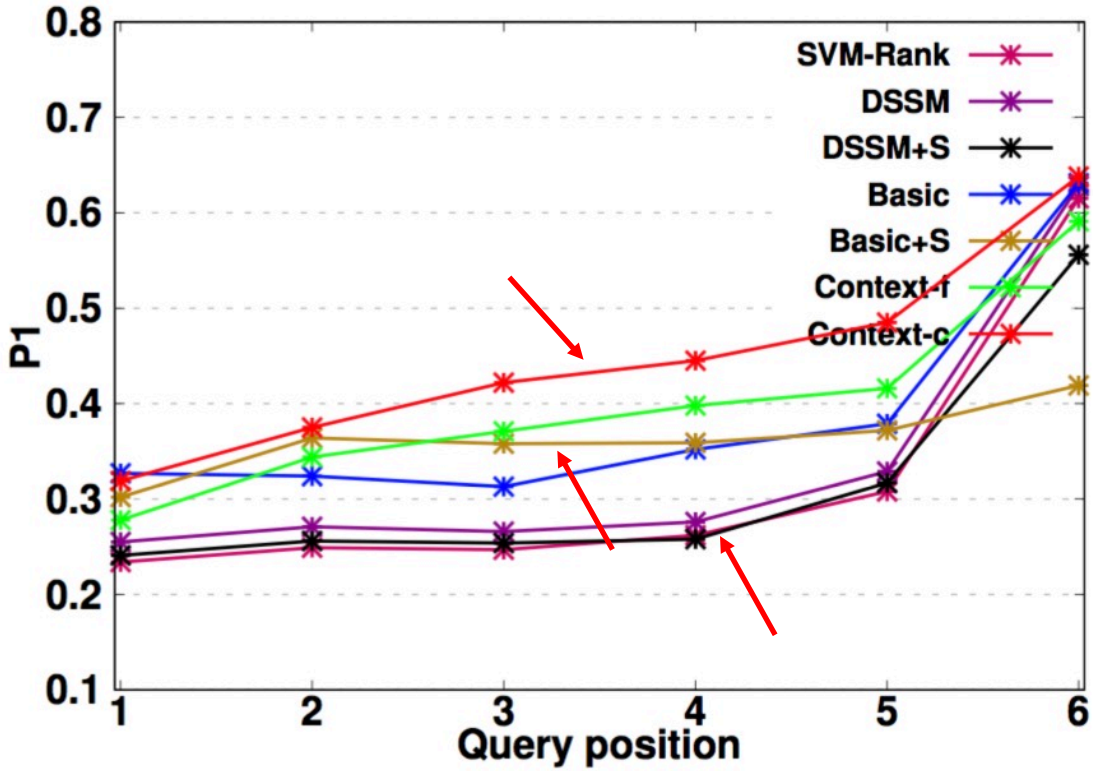
COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Context Analysis

- How does context signal help?
  - For each session length, we plot the average precision at different query position.
  - We take sessions with length of 6 as example



Observations

- Accuracies of non-context methods (SVM-Rank, DSSM, Basic) stay quite stable at different query positions.
- Simple query concatenation (DSSM+S, Basic+S) improves in the beginning, but suffers later.
- Acc. of our methods (Context-f, Context-c) consistently goes up with more queries as context.

COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Case Study

Session		Cacio : You : You : Caillou
Program		Caillou
Model	**Query**	**Example 1**
BM25	-	Pacific Rim : Now You See Me : Now You See Me : ★
SVM^{rank}	-	Pacific Rim : Now You See Me : Now You See Me : ★
DSSM	3-gram	Pacific Rim : Young : Young : ★
DSSM+S	3-gram	Pacific Rim : The Young and the Restless : The Young and the Restless : ★
Basic	char	★ (0.81) : ★ (0.80) : ★ (0.80) : ★ (1.0)
Basic	word	Child Genius (0.03) : ★ (0.57) : ★ (0.57) : ★ (1.0)
Basic	comb	Paw Patrol (0.17) : ★ (0.83) : ★ (0.83) : ★ (1.0)
Basic+S	comb	Paw Patrol (0.15) : Paw Patrol (0.25) : ★ (0.75) : ★ (0.98)
Context-f	char	Lego Ninjago (0.30) : ★ (0.79) : ★ (0.90) : ★ (0.99)
Context-f	word	Paw Patrol (0.30) : ★ (0.62) : ★ (0.98) : ★ (1.0)
Context-f	comb	Lego Ninjago (0.03) : ★ (0.60) : ★ (0.98) : ★ (1.0)
Context-c	char	★ (0.96) : ★ (0.99) : ★ (0.99) : ★ (1.0)
Context-c	word	Wallykazam (0.07) : ★ (0.59) : ★ (0.86) : ★ (1.0)
Context-c	comb	Paw Patrol (0.17) : ★ (0.93) : ★ (1.0) : ★ (1.0)

Note: the first line is a session with four queries, the second line is the ground truth label; ★ denotes the prediction is correct.

- An ambiguous session because of speech recognition(SR) errors.
- Character-level modeling is effective for SR errors.
- Basic models produce the same confidence scores for the second and third query.

COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Conclusion

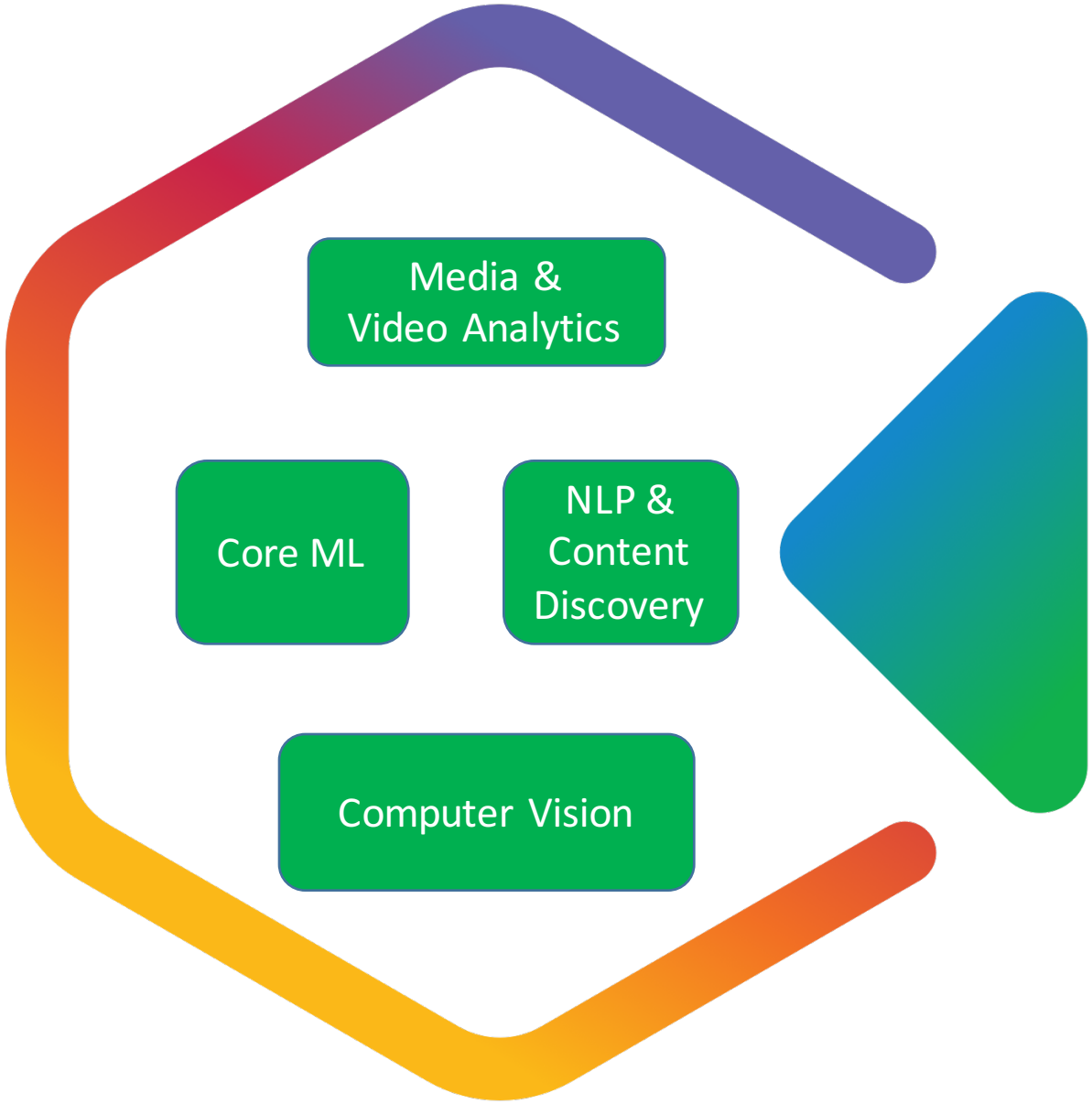
- As a first step, we study the voice navigational queries to help users find the TV programs they are looking for.
- We articulate the challenges of this task, which we tackle by combining word and character-level query representation and modeling session contexts, both using hierarchical RNN modules.
- Extensive experiments on large real datasets show our methods can effectively copy with ASR errors and ambiguity.

COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO

Comcast Applied Artificial Intelligence Group



- Video
- High Speed Internet
- Home Security / Automation
- Customer Service
- Universal Parks
- Media Properties

Thanks for listening!
Any question?

COMCAST

UNIVERSITY OF MARYLAND

UNIVERSITY OF WATERLOO