Contextual Media Retrieval Using Natural Language Queries

IMPRS-CS PhD Application Talk

Sreyasi Nag Chowdhury

Master’s Thesis Supervisors
Dr. Mario Fritz
Dr. Andreas Bulling

Adviser
M.Sc. Mateusz Malinowski
Outline

• Motivation and Overview
• Contextual Media Retrieval System
• Results and Conclusion
Motivation

“Collective Memory”
of media content

Spatio-temporal exploration of media on wearable devices
System Overview

Demonstration
System Overview

Demonstration : Spatial Exploration
System Overview

Demonstration : Temporal Exploration
System Overview

Dynamic-Egocentric environment

Natural Language Voice Query

Images and Videos

Xplore-M-Ego
Related Work

<table>
<thead>
<tr>
<th>Category</th>
<th>Existing Functions</th>
<th>Our contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Related Work

<table>
<thead>
<tr>
<th>Category</th>
<th>Existing Functions</th>
<th>Our contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatio-temporal Media Retrieval</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Related Work

<table>
<thead>
<tr>
<th>Category</th>
<th>Existing Functions</th>
<th>Our contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatio-temporal Media Retrieval</td>
<td>Browsing media collections in a static allocentric setting; Click-based GUI</td>
<td></td>
</tr>
</tbody>
</table>
## Related Work

<table>
<thead>
<tr>
<th>Category</th>
<th>Existing Functions</th>
<th>Our contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatio-temporal Media Retrieval</td>
<td>Browsing media collections in a static allocentric setting; Click-based GUI</td>
<td>Browsing media collections in a dynamic egocentric setting; hands-free GUI</td>
</tr>
</tbody>
</table>
## Related Work

<table>
<thead>
<tr>
<th>Category</th>
<th>Existing Functions</th>
<th>Our contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatio-temporal Media Retrieval</td>
<td>Browsing media collections in a static allocentric setting; Click-based GUI</td>
<td>Browsing media collections in a dynamic egocentric setting; hands-free GUI</td>
</tr>
<tr>
<td>Natural Language Query Processing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Related Work

<table>
<thead>
<tr>
<th>Category</th>
<th>Existing Functions</th>
<th>Our contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatio-temporal Media Retrieval</td>
<td>Browsing media collections in a static allocentric setting; Click-based GUI</td>
<td>Browsing media collections in a dynamic egocentric setting; hands-free GUI</td>
</tr>
<tr>
<td>Natural Language Query Processing</td>
<td>Question-answering w.r.t. a static world; Returning textual information</td>
<td></td>
</tr>
</tbody>
</table>
## Related Work

<table>
<thead>
<tr>
<th>Category</th>
<th>Existing Functions</th>
<th>Our contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatio-temporal Media Retrieval</td>
<td>Browsing media collections in a static allocentric setting; Click-based GUI</td>
<td>Browsing media collections in a dynamic egocentric setting; hands-free GUI</td>
</tr>
<tr>
<td>Natural Language Query Processing</td>
<td>Question-answering w.r.t. a static world; Returning textual information</td>
<td>Question-answering w.r.t. a dynamic world; Returning media files</td>
</tr>
</tbody>
</table>
## Related Work

<table>
<thead>
<tr>
<th>Category</th>
<th>Existing Functions</th>
<th>Our contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatio-temporal Media Retrieval</td>
<td>Browsing media collections in a static allocentric setting; Click-based GUI</td>
<td>Browsing media collections in a dynamic egocentric setting; hands-free GUI</td>
</tr>
<tr>
<td>Natural Language Query Processing</td>
<td>Question-answering w.r.t. a static world; Returning textual information</td>
<td>Question-answering w.r.t. a dynamic world; Returning media files</td>
</tr>
<tr>
<td>Media Retrieval Using Natural Language Queries</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Related Work

<table>
<thead>
<tr>
<th>Category</th>
<th>Existing Functions</th>
<th>Our contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatio-temporal Media Retrieval</td>
<td>Browsing media collections in a static allocentric setting; Click-based GUI</td>
<td>Browsing media collections in a dynamic egocentric setting; hands-free GUI</td>
</tr>
<tr>
<td>Natural Language Query Processing</td>
<td>Question-answering w.r.t. a static world; Returning textual information</td>
<td>Question-answering w.r.t. a dynamic world; Returning media files</td>
</tr>
<tr>
<td>Media Retrieval Using Natural Language Queries</td>
<td>Retrieving media based on scene contents; Using short structured phrases as queries; Does not take into account user’s context</td>
<td></td>
</tr>
</tbody>
</table>
## Related Work

<table>
<thead>
<tr>
<th>Category</th>
<th>Existing Functions</th>
<th>Our contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatio-temporal Media Retrieval</td>
<td>Browsing media collections in a static allocentric setting; Click-based GUI</td>
<td>Browsing media collections in a dynamic egocentric setting; hands-free GUI</td>
</tr>
<tr>
<td>Natural Language Query Processing</td>
<td>Question-answering w.r.t. a static world; Returning textual information</td>
<td>Question-answering w.r.t. a dynamic world; Returning media files</td>
</tr>
<tr>
<td>Media Retrieval Using Natural Language Queries</td>
<td>Retrieving media based on scene contents; Using short structured phrases as queries; Does not take into account user's context</td>
<td>Retrieving media based on geographic location; Using rich complete natural language sentences as queries; Takes into account user's context</td>
</tr>
</tbody>
</table>
Outline

• Motivation and Overview

• Contextual Media Retrieval System

• Results and Conclusion
Google Glass Client

1. App started with voice command: "Ok Glass Explore Nearby"
2. Voice query detected and transcribed to text
3. GPS location recorded by location sensor; Viewing direction recorded by orientation sensor
4. sendQuery(query, metadata)
5. Dynamic database created; query modified
6. Query parsed by semantic parser; answers predicted from denotations
7. Media files retrieved from collective memory
8. receiveResult(media file)

Python Server

0. TCP socket connection established
1. Dedicated port listens for client connections indefinitely

Sreyasi Nag Chowdhury | Contextual Media Retrieval Using Natural Language Queries
Google Glass Client

1. App started with voice command: "Ok Glass Explore Nearby"

2. Voice query detected and transcribed to text

3. GPS location recorded by location sensor; Viewing direction recorded by orientation sensor

Python Server

0. TCP socket connection established

1. Dedicated port listens for client connections indefinitely

2. Dynamic database created; query modified

3. Query parsed by semantic parser; answers predicted from denotations

4. sendQuery(query, metadata)

5. receiveResult(media file)

6. Media files retrieved from collective memory

Returned results viewed by user
Google Glass Client

1. App started with voice command: "Ok Glass Explore Nearby"
2. Voice query detected and transcribed to text
3. GPS location recorded by location sensor; Viewing direction recorded by orientation sensor
4. sendQuery(query, metadata)
5. Returned results viewed by user

Python Server

0. TCP socket connection established
1. Dedicated port listens for client connections indefinitely
2. Dynamic database created; query modified
3. Query parsed by semantic parser; answers predicted from denotations
4. Media files retrieved from collective memory
5. receiveResult(media file)
Google Glass Client

1. App started with voice command: "Ok Glass Explore Nearby"
2. Voice query detected and transcribed to text
3. GPS location recorded by location sensor; Viewing direction recorded by orientation sensor
4. sendQuery(query, metadata)

Python Server

0. TCP socket connection established
1. Dedicated port listens for client connections indefinitely
2. Dynamic database created; query modified
3. Query parsed by semantic parser; answers predicted from denotations
4. Media files retrieved from collective memory
5. receiveResult(media file)
Google Glass Client

1. App started with voice command: "Ok Glass Explore Nearby"

2. Voice query detected and transcribed to text

3. GPS location recorded by location sensor; Viewing direction recorded by orientation sensor

4. sendQuery(query, metadata)

Python Server

5. TCP socket connection established

6. Dedicated port listens for client connections indefinitely

7. Dynamic database created; query modified

8. Query parsed by semantic parser; answers predicted from denotations

7. Media files retrieved from collective memory
Google Glass Client

1. App started with voice command: "Ok Glass Explore Nearby"

2. Voice query detected and transcribed to text

3. GPS location recorded by location sensor; Viewing direction recorded by orientation sensor

4. sendQuery(query, metadata)

Python Server

0. TCP socket connection established

1. Dedicated port listens for client connections indefinitely

2. Dynamic database created; query modified

3. Query parsed by semantic parser; answers predicted from denotations

4. Media files retrieved from collective memory

5. receiveResult(media file)
Google Glass Client

1. App started with voice command: "Ok Glass Explore Nearby"
2. Voice query detected and transcribed to text
3. GPS location recorded by location sensor; Viewing direction recorded by orientation sensor
4. sendQuery(query, metadata)
5. TCP socket connection established
6. Dedicated port listens for client connections indefinitely
7. Dynamic database created; query modified
8. Query parsed by semantic parser; answers predicted from denotations
9. Media files retrieved from collective memory
10. Returned results viewed by user
11. receiveResult(media file)
**Question – Answering**

Q&A Model of Percy Liang

("Learning Dependency-based Compositional Semantics", Liang et al.)
Question – Answering

Q&A Model of Percy Liang
("Learning Dependency-based Compositional Semantics", Liang et al.)
**Question – Answering**

(x) \(\rightarrow\) Semantic Parsing \(\rightarrow\) Evaluation \(\rightarrow\) (y)

\(\theta\) (parameters) \(\rightarrow\) Evaluation \(\rightarrow\) (y)

\(w\) (world) \(\rightarrow\) Evaluation \(\rightarrow\) (y)

(state) \(\rightarrow\) (logical form) \(\rightarrow\) (answer)

*state with the largest area*

\(z \sim p_\theta(z \mid x)\)

\(y = [z]_w\)

**Constraint Satisfaction Problem**

**Q&A Model of Percy Liang**

(“Learning Dependency-based Compositional Semantics”, Liang et al.)
Question – Answering

Q&A Model of Percy Liang
(“Learning Dependency-based Compositional Semantics”, Liang et al.)
Unambiguous query

State with the largest area

Single correct answer

Alaska

Q&A Model of Percy Liang
("Learning Dependency-based Compositional Semantics", Liang et al.)
Question – Answering

Q&A Model of Percy Liang
(“Learning Dependency-based Compositional Semantics”, Liang et al.)
Q&A Model of Percy Liang
(“Learning Dependency-based Compositional Semantics”, Liang et al.)
Our Q&A Model

 Ambiguous query

*what is there in front of MPI-INF?*
Question – Answering

Our Q&A Model

Ambiguous query

Subjective; multiple correct answers

what is there in front of MPI-INF?

$z \sim p_{\theta}(z \mid x)$

$y = [z]_w$
Our Q&A Model

what is there in front of MPI-INF?

Ambiguous query

Dynamic world

Subjective; multiple correct answers
Question – Answering

Our Q&A Model

what is there in front of MPI-INF?

Ambiguous query

Dynamic world

Media as answer

Subjective; multiple correct answers

$z \sim p_\theta(z | x)$

$y = [z]_w$
Dynamic-Egocentric Extension

World (w)
Dynamic-Egocentric Extension

World (w)
**Dynamic-Egocentric Extension**

![Diagram showing the extension of the World (w) into Static World (w_s).](image)

- **World (w)**
  - **Static World (w_s)**
    - cafe('mensa', 49.2560, 7.0454).
    - building('mpi_inf', 49.2578, 7.0460).
Dynamic-Egocentric Extension

World (w)

Static World (w_s)
- cafe('mensa', 49.2560, 7.0454).
- building('mpi_inf', 49.2578, 7.0460).

Dynamic World (w_d)
Dynamic-Egocentric Extension

\[ \text{World} \ (w) \]

\[ \text{Static World} \ (w_s) \]
- cafe('mensa', 49.2560, 7.0454).
- building('mpi_inf', 49.2578, 7.0460).

\[ \text{Dynamic World} \ (w_d) \]
Dynamic-Egocentric Extension

World \((w)\)

Static World \((w_s)\)
- cafe('mensa', 49.2560, 7.0454).
- building('mpi_inf', 49.2578, 7.0460).

Dynamic World \((w_d)\)

User Metadata \((w_{du})\)
- person(49.2578, 7.0460, 'n').
- day(20150220).
Dynamic-Egocentric Extension

**World** ($w$)

**Static World** ($w_s$)
- cafe('mensa', 49.2560, 7.0454).
- building('mpi_inf', 49.2578, 7.0460).

**User Metadata** ($w_{du}$)
- person(49.2578, 7.0460, 'n').
- day(20150220).

**Dynamic World** ($w_d$)

**Collective Memory** ($w_{dm}$)
- image('img_20141111_165828', 20141111, 1, 49.2566, 7.0442, 'november').
- video('vid_20141121_120149', 20141121, 49.2569, 7.0456, 'november').
Qualitative Results

“What building is to the left of MPI-SWS?”
Qualitative Results

“What building is to the left of MPI-SWS?”
“What building is to the left of MPI-SWS?”
Qualitative Results

“What building is to the left of MPI-SWS?”
Qualitative Results

“What building is to the left of MPI-SWS?”

“What is near MPI-INF?”
Qualitative Results

“What building is to the left of MPI-SWS?”

“What is near MPI-INF?”

“What did this place (MPI-INF) look like in December?”
Outline

• Motivation and Overview

• Contextual Media Retrieval System

• Results and Conclusion
Evaluation

Agreement and Disagreement between users

* Model tested on 500 test queries
Evaluation

Agreement and Disagreement between users

* Model tested on 500 test queries

Total agreement 26.67%
Agreement and Disagreement between users

Majority agreement

~ 40%

* Model tested on 500 test queries
Evaluation

Agreement and Disagreement between users

Disagreement ~ 25%

* Model tested on 500 test queries
Evaluation

Agreement and Disagreement between users

Future scope for personalization

~ 25%

* Model tested on 500 test queries
Evaluation

Study of human reference frame resolution
Evaluation

Study of human reference frame resolution

Future scope for using other Knowledgebases
Summary

We have:

• Instantiated a "Collective Memory" of media content

• Developed a novel architecture for media retrieval with natural language voice queries in a dynamic setting - Xplore-M-Ego

• Integrated ‘egocentrism’ to media retrieval
Summary

We have:

• Instantiated a “Collective Memory” of media content

• Developed a novel architecture for media retrieval with natural language voice queries in a dynamic setting - *Xplore-M-Ego*

• Integrated ‘egocentrism’ to media retrieval

Thank You
References

• Photo Tourism: Exploring Photo Collections in 3D
  Noah Snavely, Steven M. Seitz, Richard Szeliski

• Video Collections in Panoramic Contexts
  J.Tompkin, F.Pece, R.Shah, S.Izadi, J.Kautz, C.Theobalt

• Videoscapes: Exploring Sparse, Unstructures Video Collections
  J.Tompkin, K. In Kim, J.Kautz, C.Theobalt

• PhotoScope: Visualizing Spatiotemporal Coverage of Photos for Construction Management
  F.Wu, M.Tory
References

• Learning Dependency-Based Compositional Semantics
  Percy Liang, Michael I. Jordan, Dan Klein

• A multi-world approach to question answering about real-world scenes based on uncertain input
  M. Malinowski, M. Fritz

• Image Retrieval with Structured Object Queries Using Latent Ranking SVM
  T.Lan, W.Yang, Y.Wang, G.Mori

• Interpretation of Spatial Language in a Map Navigation Task
  M. Levit, D. Roy
Extra Material
Contribution

• Instantiation of a “Collective Memory” of media files

• Extension of question-answering to a dynamic setting

• Extension of spatio-temporal exploration of media to a dynamic setting

• Incorporation of ‘egocentrism’ to media retrieval

• Use of natural language voice queries for media retrieval
System Overview

Modules of \textit{Xplore-M-Ego}

- \textbf{The Google Glass:} User Interface
- \textbf{Pre-processing:} Modification of query, Mapping of a dynamic environment to a static environment
- \textbf{Semantic Parser + Denotation:} Semantic parsing and prediction of answer
- \textbf{Collective Memory:} Store of media files
## Related Work

- Spatio-temporal Media Retrieval

<table>
<thead>
<tr>
<th>Paper</th>
<th>Author(s)</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photo tourism: exploring photo collections in 3D</td>
<td>N. Snavely, S. M. Seitz, and R. Szeliski</td>
<td>Exploration of popular world sites by browsing through images</td>
</tr>
</tbody>
</table>
### Related Work

**Natural Language Question-Answering**

<table>
<thead>
<tr>
<th>Paper</th>
<th>Author(s)</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Dependency-based compositional Semantics</td>
<td>P. Liang, M. I. Jordan, and D. Klein</td>
<td>Training of a semantic parser with question-answer pairs; single static world approach</td>
</tr>
<tr>
<td>A multi-world approach to question answering about real-world scenes based on uncertain input</td>
<td>M. Malinowski and M. Fritz</td>
<td>Question-answering task based on real world indoor images; static multi-world approach</td>
</tr>
</tbody>
</table>
Related Work

- Media Retrieval with Natural Language Queries

<table>
<thead>
<tr>
<th>Paper</th>
<th>Author(s)</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Towards surveillance video search by</td>
<td>S. Tellex and D. Roy</td>
<td>Retrieval of video frames from surveillance videos with spatial relations “across” and “along”</td>
</tr>
<tr>
<td>natural language query</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image retrieval with structured object</td>
<td>T. Lan, W. Yang, Y. Wang, and G.</td>
<td>Retrieval of images based on scene contents using short structured phrases as queries</td>
</tr>
<tr>
<td>queries using latent ranking SVM</td>
<td>Mori</td>
<td></td>
</tr>
</tbody>
</table>
Data Collection

1. Map information: OpenStreetMap

Contains –
- Type of the entity
- GPS coordinates
- Name
- Address
Data Collection

2. Collection of media files: *Collective Memory*

** Media files were captured with smart phones
3. Training and Test data

- **Synthetically-generated Data**
  - “What is there in front of MPI-INF?”, answer(A, (frontOf(A, ‘mpi inf’))))
  - “What is there behind MPI-INF?”, answer(A, (behind(A, ‘mpi inf’))))
  - “What is there on the right of MPI-INF?”, answer(A, (rightOf(A, ‘mpi inf’))))
  - “What is there on the left of MPI-INF?”, answer(A, (leftOf(A, ‘mpi inf’))))

- **Real-world Data**
  - “What is there on the left of MPI-INF?”, ‘img_20141102_123406’
  - “What is on the left of MPI-INF?”, ‘img_20141113_160930’
  - “What is to the left of MPI-INF?”, ‘img_20141109_134914’
  - “What is on the left side of MPI-INF?”, ‘img_20141115_100705’
Data Collection

“What is there beside MPI-INF?”

“What is on the left of E 1.3?”

“What is in front of the campus center?”

“How does the campus bus stop look?”

“What is there on the right side of the university campus?”

“What is in front of the university bus terminal?”
Semantic Parser

Dependency-based Compositional Semantics (DCS) by Percy Liang

Example: *major city in California*

\[
z = \langle \text{city}; 1: \langle \text{major} \rangle; 1: \langle \text{loc}; \frac{2}{1}: \langle \text{CA} \rangle \rangle
\]

- DCS tree defines relations between predicates
- Denotation are solutions satisfying the relations
- *city*, *major*, *loc*, *CA* are predicates

(a) DCS tree

(b) Lambda calculus formula

\[
\lambda c \exists m \exists \ell \exists s .
\]

\[
\text{city}(c) \land \text{major}(m) \land \text{loc}(\ell) \land \text{CA}(s) \land c_1 = m_1 \land c_1 = \ell_1 \land \ell_2 = s_1
\]

(c) Denotation: \([z]_w = \{ \text{SF, LA, \ldots} \}\)
World($w$):

state('california','ca', 'sacramento', 23.67e+6, 158.0e+3,31, 'los angeles', 'san diego', 'san francisco', 'san jose').

city('alabama','al','birmingham',284413).

river('arkansas',2333,['colorado','kansas', 'oklahoma','arkansas']).

mountain('alaska','ak','mckinley',6194).

road('86',['massachusetts','connecticut']).
country('usa',307890000,9826675).

Example Questions

“What is the highest point in Florida?”

“Which State has the shortest river?”

“What is the capital of Maine?”

“What are the populations of states through which the Mississippi river run?”

“Name all the lakes of US?”
Semantic Parser

Learning in DCS

Objective:

$$\max_\theta \sum_z p(y \mid z, w) \ p(z \mid x, \theta)$$

Interpretation Semantic parsing

EM-like Algorithm:

parameters $\theta$

(0.3, −1.4, . . . , 0.6)

$k$-best list

enumerate/score DCS trees

numerical optimization (L-BFGS)

tree3 ✓
tree8 ✓
tree2 ✗
tree4 ✗
tree9 ✗
**Semantic Parser**

- Induction of logical forms

  - Logical forms (DCS trees) induced as latent variables according to a probability distribution parametrized with $\theta$

  - Answer $y$ evaluated with respect to world $w$

  \[
  z \sim p_\theta(z \mid x) \\
  y = \arg\max_w [z]_w
  \]
• Induction of logical forms

Requirements –

A set of rules/predicates:

city(cityid(City,St)) :- state(State,St,_,_,_,_,_,_,City,_,_,_).
loc(cityid(City,St),stateid(State)) :- state(State,St,_,_,_,_,_,City,_,_,_).
river(riverid(R)) :- river(R,_,_).
loc(cityid(City,St),stateid(State)) :- city(State,St,City,).
traverse(riverid(R),stateid(S)) :- river(R,_,States), member(S,States).
area(stateid(X),squared mile(Area)) :- state(X,_,_,_,Area,_,_,_,_,_,_).
population(countryid(X),Pop) :- country(X,Pop,_,).
major(X) :- city(X), population(X,moreThan(150000)).
• Induction of logical forms

Requirements –

A set of lexical triggers($L$):

$$<(\text{function words}; \text{predicate})>$$

- (most, size).
- (total, sum).
- (called, nameObj).

$$<(\text{POS tags}; [\text{predicates}])>$$

- (WRB; loc)
- ([NN;NNS]; [city, state, country, lake, mountain, river, place])
- ([NN;NNS]; [person, capital, population])
- ([NN;NNS; JJ]; [len, negLen, size, negSize, elevation])
- ([NN;NNS; JJ]; [negElevation, density, negDensity, area, negArea])
- (JJ; major)

Augmented Lexicon($L+$):

- (long, len).
- (large, size).
- (small, negSize).
- (high, elevation).
Media Retrieval from Denotations

World(w):

image(`img_20141111_165828',20141111,49.2566,7.0442,'november').
video(`vid_20141121_120149',20141121,49.2569,7.0456,'november').
cafe(`mensa',49.2560,7.0454).
building(`mpi_inf',49.2578,7.0460).
bank(`postbank',49.2556,7.0449).

Example Questions

“What is there on the right of MPI-INF?”
“What is there in front of postbank?”
“What is there on the left of Mensa?”
“What is there near Science Park?”
“What happened here one day ago?”
“What does this place look like in December?”
Lexical triggers:

<table>
<thead>
<tr>
<th>Basic lexicon $L$</th>
<th>Augmented lexicon $L+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>([WP,WDT], [image,video]). (NN, atm,building,cafe,highway,parking,research_institution, restaurant,shop,sport,tourism,university]). (JJS, [nearest]). ([NN,NNS,VB], [view]). (VBD, [view]).</td>
<td>(front, frontOf). (behind, behind). (right, rightOf). (left, leftOf).</td>
</tr>
</tbody>
</table>

Prediction accuracy: 17.9%                               Prediction accuracy: 47%
Dynamic-Egocentric Extension

Static Database of Geographic Facts $w_s$

```
atm('postbank_atm',49.2573855,7.0430358,49.2574,7.0430).
bank('bank1saar',49.2545957,7.0401859,49.2546,7.0402).
bar('canossa',49.2572934,7.0429204,49.2573,7.0429).
building('department_of_culture',49.25343,7.0414877,49.2534,7.0415).
cafe('icoffee',49.2574952,7.0453556,49.2575,7.0454).
highway('campus',49.25573,7.0389795,49.2557,7.0390).
library('state_library',49.253353,7.038327,49.2534,7.0383).
parking('uni_nord',49.25751,7.041421,49.2575,7.0414).
research_institution('dfki',49.25717,7.041499,49.2572,7.0415).
```
Dynamic-Egocentric Extension

Dynamic Database of User Metadata $w_{du}$

person(49.2578,7.0454,‘n’).
day(20141104).

Dynamic Database of Media Content $w_{dm}$

image(‘img_20141111_165828’,20141111,49.2566,7.0442,‘november’).
image(‘img_20141112_092045’,20141112,49.2554,7.0396,‘november’).
video(‘vid_20141121_120149’,20141121,49.2569,7.0456,‘november’).
video(‘vid_20141123_165241’,20141123,49.2530,7.0338,‘november’).
POS tags from Penn Treebank

- WRB : Wh-adverb
- NN : Noun, singular or mass
- NNS : Noun, plural
- JJ : Adjective
- WP : Wh-pronoun
- WDT : Wh-determiner
- NN : Noun, singular or mass
- JJS : Adjective, superlative
- NNS : Noun, plural
- VB : Verb
- VBD : Verb, past tense
Reason behind hard-coding spatial relations

- What is there left/VBN of MPI?
- What is there on the left/NN of MPI?
- What is there in front/NN of MPI?
- What is there behind/IN MPI?
- What is there right/RB of MPI?
- What is there on the right/NN of MPI?
# Predicates used in Xplore-M-Ego

Table 4.1: Definitions of predicates in our DCS

<table>
<thead>
<tr>
<th>Predicates</th>
<th>Definitions</th>
<th>Example Query</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spatial</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>frontOf(A,B)</td>
<td>lat(B)&gt;lat(A), lon(A)=lon(B)</td>
<td>“what is in front of A?”</td>
</tr>
<tr>
<td>behind(A,B)</td>
<td>lat(B)&lt;lat(A), lon(A)=lon(B)</td>
<td>“what is behind A?”</td>
</tr>
<tr>
<td>rightOf(A,B)</td>
<td>lon(B)&gt;lon(A), lat(A)=lat(B)</td>
<td>“what is on the right of A?”</td>
</tr>
<tr>
<td>leftOf(A,B)</td>
<td>lon(B)&lt;lon(A), lat(A)=lat(B)</td>
<td>“what is on the left of A?”</td>
</tr>
<tr>
<td><strong>Temporal</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>view2(M,B)</td>
<td>month(B)=M, lat(B)=user’s lat, lon(B)=user’s lon</td>
<td>“how did this place look in M?”</td>
</tr>
<tr>
<td>view1(B)</td>
<td>timestamp(B)=user’s timestamp, lat(B)=user’s lat, lon(B)=user’s lon</td>
<td>“what happened here 5 days ago?”</td>
</tr>
</tbody>
</table>

Here, B is a media file. A is a geographical entity (e.g. ‘MPI’) and M is a month (e.g. ‘May’) uttered as part of the query; ‘lat’ and ‘lon’ stand for GPS latitude and longitude; day and person are predicates in \( w_{du} \)
Results and Evaluation

- Synthetically generated question-answer pairs used for training and testing
- Maximum prediction accuracy – 47%
Results and Evaluation

Performance Measures:

- $q_m = \text{number of queries with media retrievals}$
- $q_r = \text{number of queries with relevant retrievals among } q_m$
- $q_t = \text{number of queries with textual retrievals and no retrievals}$

- average precision $= \frac{q_r}{q_m}$
- average recall $= \frac{q_r}{q_m + q_t}$
Results and Evaluation

• “human-in-the-loop” training of the model
  
  o Five different models were trained
  
  o Training accuracies ranged from 42.6% to 48.8%
  
  o The best model based on training accuracy was used for further evaluations
Results and Evaluation

• “human-in-the-loop” training of the model

- It is a method of training the semantic parser by human users through relevance feedback

- “Correct”/“Wrong” decisions are made solely based on the predicted answers

- The models are trained with real questions from human users
Results and Evaluation

• “human-in-the-loop” training of the model
  
  o Automatic training of the semantic parser with the real data was not possible because –
    
    • GPS coordinates of media files showing a particular entity does not match that of the map data
  
    • Humans are inconsistent with regards to reference frames
  
    • Question-answer pairs didn’t follow any pattern
  
    • Denotations (often more than one answer) never matched with true answers, hence EM-like algorithm failed to learn
Results and Evaluation

Human evaluation of model trained with real-world data

- RealModel - model trained with real-world data
- Relevance feedback collected from five users
- Overall percentage of relevant retrievals = 26.67%
Results and Evaluation

- Recall of SynthModel = 15.88%
- Recall of RealModel = 26.67%
Evaluation

Human evaluation of temporal and contextual Q&A

- Five hypothetical locations and viewing directions provided to users
- Relevance feedback collected for retrievals following a canonical reference frame and a user-centric reference frame
Evaluation

Human evaluation of temporal and contextual Q&A

Relevance using Canonical and User-centric Reference Frame

- Canonical and User-centric reference frame:
  - "front of"
  - "left of"
  - "right of"
  - "behind"

Original: What is there in front of MPI-INF?

Altered: What is there on the right of MPI-INF?
Discussion

Problem with matching GPS coordinates

What is in front of MPI-INF?

ground-truth media

MPI-INF

retrieved media

“Front of MPI-INF”

iCoffee

What is in front of MPI-INF?
## Discussion

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Converting a dynamic world to a static world</td>
<td>Spatial and temporal references not identified</td>
</tr>
<tr>
<td>Integrating ‘egocentrism’</td>
<td>Words tagged with incorrect POS tags</td>
</tr>
<tr>
<td>Handling temporal queries</td>
<td>Arguments not identified from sentences</td>
</tr>
<tr>
<td>Collection of data</td>
<td>Scalability</td>
</tr>
<tr>
<td>Increasing the coverage of the static database</td>
<td>Reference resolution is not handled</td>
</tr>
</tbody>
</table>
Discussion

Accuracy of Performance

- Matching the exact GPS coordinates for retrievals proved to be a failure

- It was handled by rough localization by rounding the GPS coordinates to the first 6 significant digits (49.2578401 -> 49.2578)

Failure case:

<table>
<thead>
<tr>
<th>Query</th>
<th>Retrieved Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>“What does MPI-SWS look like?”</td>
<td><img src="image1.jpg" alt="Retrieved Images" /></td>
</tr>
</tbody>
</table>
Discussion

Future Work

- Integration of image processing and computer vision methods for scene understanding (similar to Malinowski et al.)

- Development of a better semantic parser in light of our discussions about its limitations

- Development of more robust location sensors in devices used for capturing media

- Generation of a consensus about reference frames for applications involving the use of spatial relations
### Summary of Quantitative Results

Table 6.2: Use of Lexicons $L$ and $L+$

<table>
<thead>
<tr>
<th></th>
<th>Untrained Model</th>
<th>Trained Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Lexicon $L$</td>
<td>6%</td>
<td>17.9%</td>
</tr>
<tr>
<td>Augmented Lexicon $L+$</td>
<td>11.23%</td>
<td>47%</td>
</tr>
</tbody>
</table>

Table 6.3: Average Precision and Average Recall of semantic parser models

<table>
<thead>
<tr>
<th></th>
<th>Average Precision</th>
<th>Average Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>SynthModel</td>
<td>50.2%</td>
<td>16%</td>
</tr>
<tr>
<td>RealModel</td>
<td>37.38%</td>
<td>28%</td>
</tr>
</tbody>
</table>

Table 6.4: Relevance feedback using different reference frames

<table>
<thead>
<tr>
<th></th>
<th>Canonical</th>
<th>User-centric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>56%</td>
<td>49.6%</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>15%</td>
<td>10.4%</td>
</tr>
</tbody>
</table>