Unsupervised Learning and Modeling of Knowledge and Intent for Spoken Dialogue Systems

YUN-NUNG (VIVIAN) CHEN HTTP://VIVIANCHEN.IDV.TW

COMMITTEE: ALEXANDER I. RUDNICKY (CHAIR)
ANATOLE GERSHMAN (CO-CHAIR)
ALAN W BLACK
DILEK HAKKANI-TÜR (Microsoft Research)

Carnegie Mellon

2015 DECEMBER

Outline

- Introduction
- Ontology Induction [ASRU'13, SLT'14a]
- Structure Learning [NAACL-HLT'15]
- Surface Form Derivation [SLT'14b]
- Semantic Decoding [ACL-IJCNLP'15]
- Intent Prediction [SLT'14c, ICMI'15]
- SLU in Human-Human Conversations [ASRU'15]
- Conclusions & Future Work

Knowledge Acquisition

SLU Modeling

Outline

Introduction

Ontology Induction [ASRU'13, SLT'14a]

Structure Learning [NAACL-HLT'15]

Surface Form Derivation [SLT'14b]

Semantic Decoding [ACL-IJCNLP'15]

Intent Prediction [SLT'14c, ICMI'15]

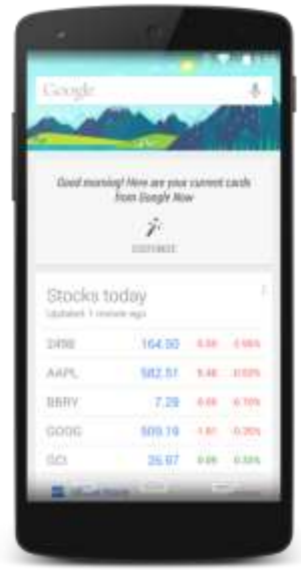
SLU in Human-Human Conversations [ASRU'15]

Conclusions & Future Work

Intelligent Assistants

Apple Siri

(2011)



Google Now

(2012)

Microsoft Cortana

(2014)

Amazon Alexa/Echo

(2014)

Facebook M

(2015)

https://www.apple.com/ios/siri/
https://www.google.com/landing/now/
http://www.windowsphone.com/en-us/how-to/wp8/cortana/meet-cortana
http://www.amazon.com/oc/echo/

Large Smart Device Population

Global Digital Statistics (2015 January)

Global Population

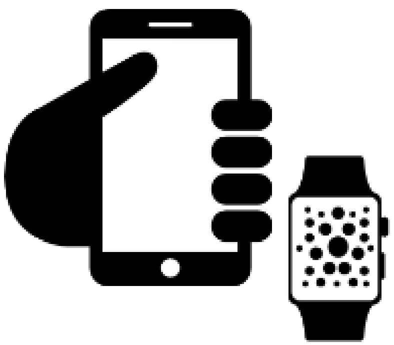
7.21B

Active Internet Users

3.01B

Active Social Media Accounts

2.08B



Active Unique Mobile Users

3.65B

The more **natural** and **convenient** input of the devices evolves towards speech.

Spoken Dialogue System (SDS)

Spoken dialogue systems are intelligent agents that are able to help users finish tasks more efficiently via spoken interactions.

Spoken dialogue systems are being incorporated into various devices (smart-phones, smart TVs, in-car navigating system, etc).

JARVIS – Iron Man's Personal Assistant

Baymax – Personal Healthcare Companion

Good SDSs assist users to organize and access information conveniently.

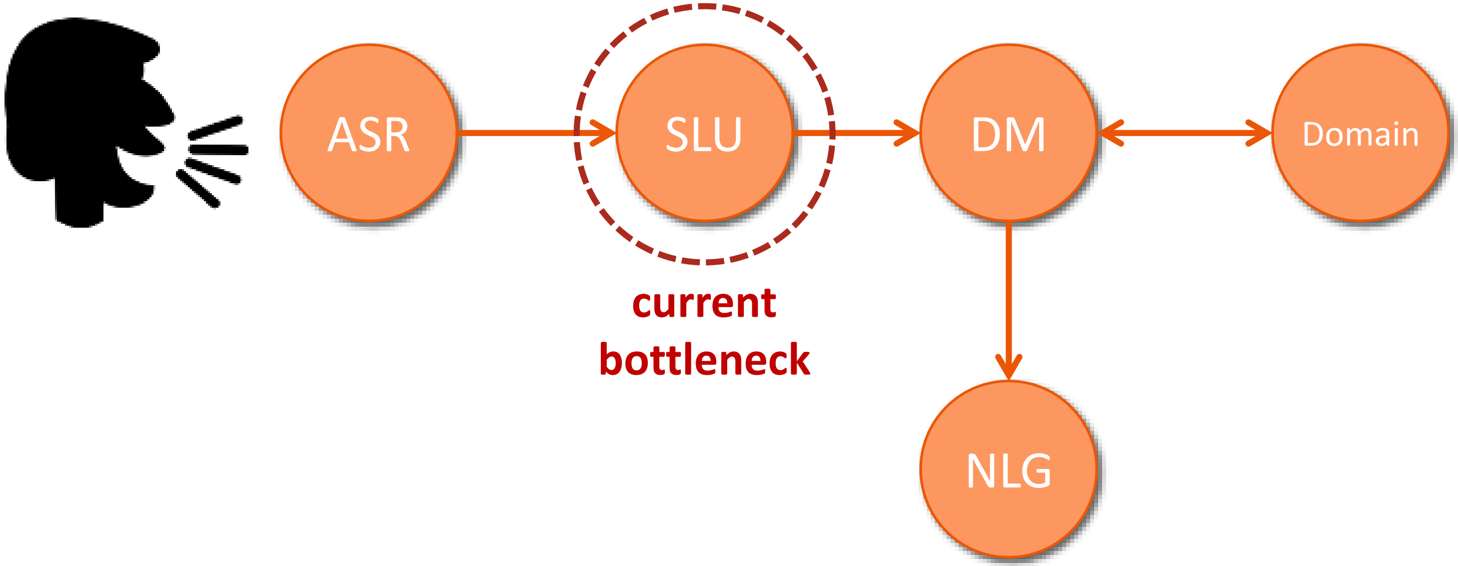
SDS Architecture

ASR: Automatic Speech Recognition

SLU: Spoken Language Understanding

DM: Dialogue Management

NLG: Natural Language Generation

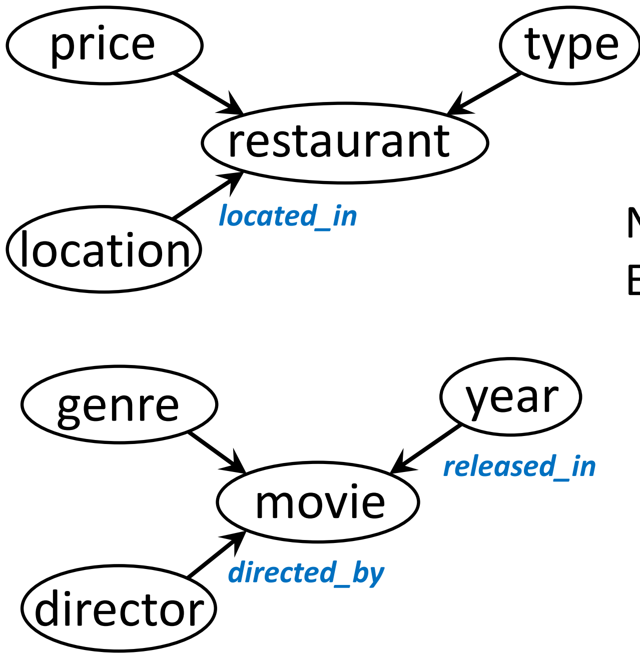


Knowledge Representation/Ontology

Traditional SDSs require **manual annotations** for **specific domains** to represent domain knowledge.

Restaurant Domain

Movie Domain



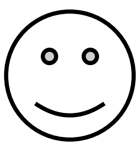
Node: semantic concept/slot
Edge: relation between concepts

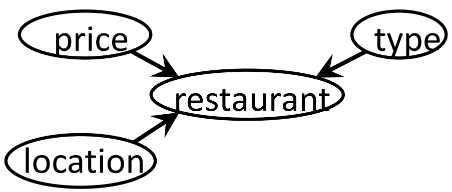
Utterance Semantic Representation

An SLU model requires a domain ontology to decode utterances into semantic forms, which contain **core content (a set of slots and slot-fillers)** of the utterance.

Restaurant Domain

find a cheap taiwanese restaurant in oakland

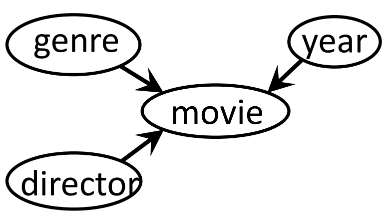




target="restaurant", price="cheap", type="taiwanese", location="oakland"

Movie Domain

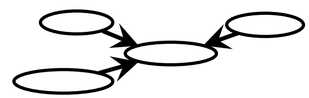
show me action movies directed by james cameron



target="movie", genre="action", director="james cameron"

Challenges for SDS

An SDS in a new domain requires

1) A hand-crafted domain ontology 
2) Utterances labelled with semantic representations
3) An SLU component for mapping utterances into semantic representations

Prior Focus

find a cheap eating place for asian food → seeking="find" target="eating place" price="cheap" food="asian food"

Manual work results in **high cost**, **long duration** and **poor scalability** of system development.

The goal is to enable an SDS to
1) automatically infer domain knowledge and then to
2) create the data for SLU modeling
in order to handle the open-domain requests.

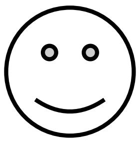
→ **fully unsupervised**

Questions to Address

1) Given unlabelled conversations, how can a system automatically induce and organize domain-specific concepts?
2) With the automatically acquired knowledge, how can a system understand utterance semantics and user intents?

Interaction Example

User



find a cheap eating place for asian food

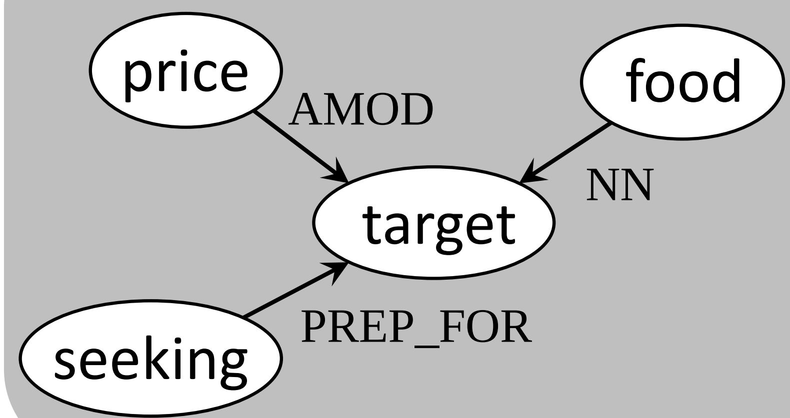
Cheap Asian eating places include Rose Tea Cafe, Little Asia, etc. What do you want to choose? I can help you go there. (navigation)

Intelligent Agent

Q: How does a dialogue system process this request?

Process Pipeline

User



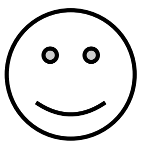
SELECT restaurant {
  restaurant.price="cheap"
  restaurant.food="asian food"
}

Predicted intent: navigation

Required Domain-Specific Information

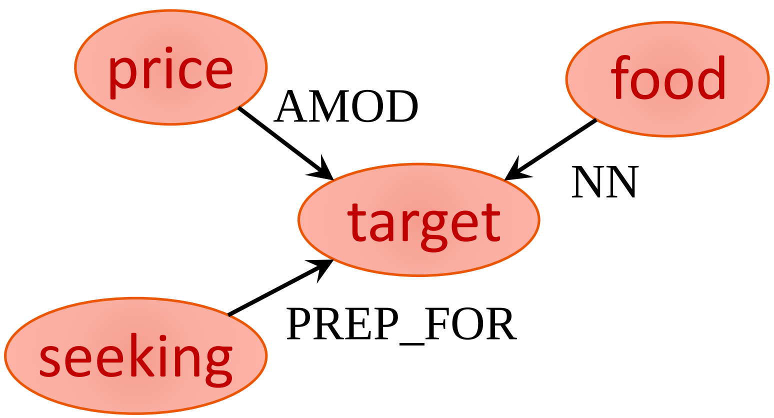
Thesis Contributions

User



find a cheap eating place for asian food

Ontology Induction *(semantic slot)*



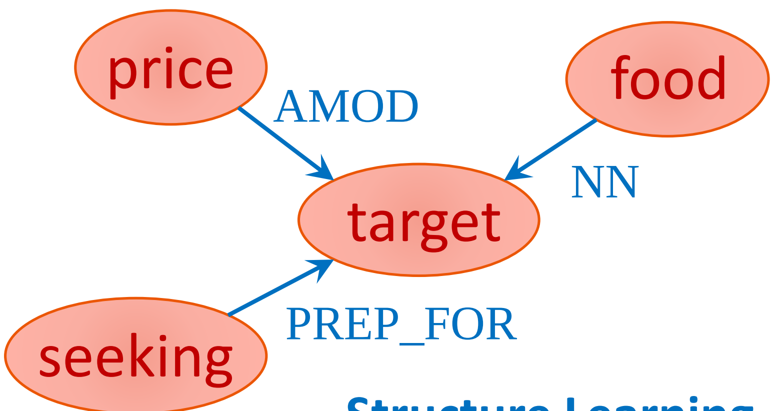
SELECT restaurant {
  restaurant.price="cheap"
  restaurant.food="asian food"
}

Predicted intent: navigation

Thesis Contributions

User

Ontology Induction *(semantic slot)*



Structure Learning

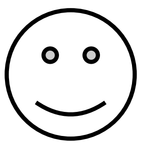
(inter-slot relation)

SELECT restaurant {
    restaurant.price="cheap"
    restaurant.food="asian food"
}

Predicted intent: navigation

Thesis Contributions

User

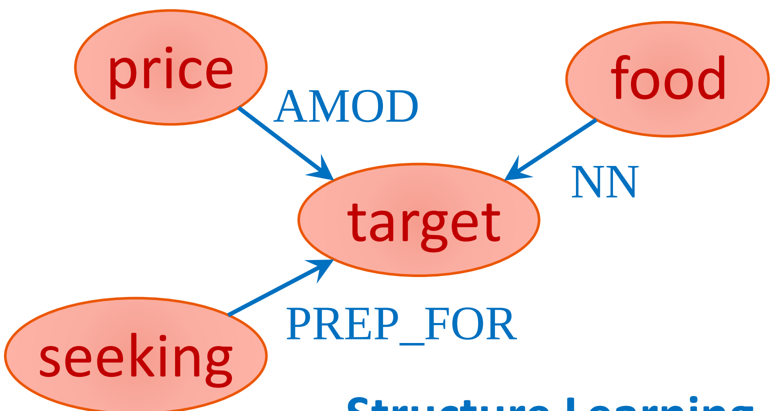


find a cheap eating place for asian food

Surface Form Derivation

(natural language)

Ontology Induction *(semantic slot)*



Structure Learning

(inter-slot relation)

SELECT restaurant {
restaurant.price="cheap"
restaurant.food="asian food"
}

Predicted intent: navigation

Thesis Contributions

User

find a cheap eating place for asian food

Surface Form Derivation

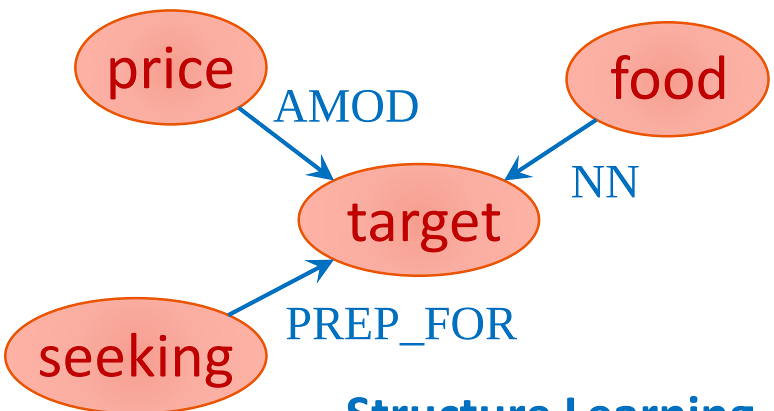
(natural language)

SELECT restaurant {
restaurant.price="cheap"
restaurant.food="asian food"
}

Semantic Decoding

Predicted intent: navigation

Ontology Induction *(semantic slot)*



Structure Learning

(inter-slot relation)

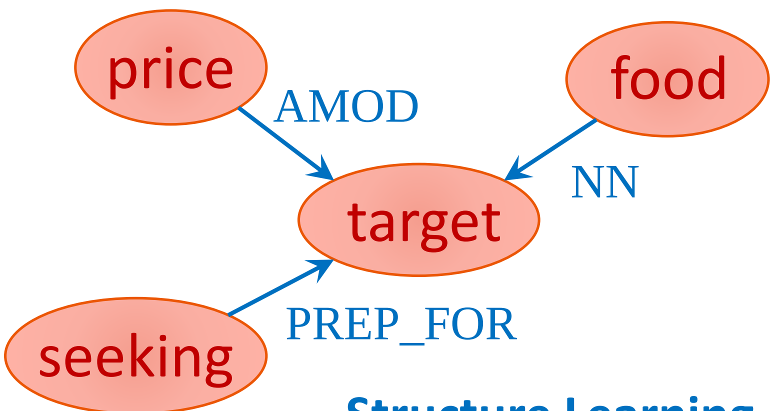
Thesis Contributions

User

find a cheap eating place for asian food

Surface Form Derivation *(natural language)*

Ontology Induction *(semantic slot)*



Structure Learning

(inter-slot relation)

SELECT restaurant {
restaurant.price="cheap"
restaurant.food="asian food"
}

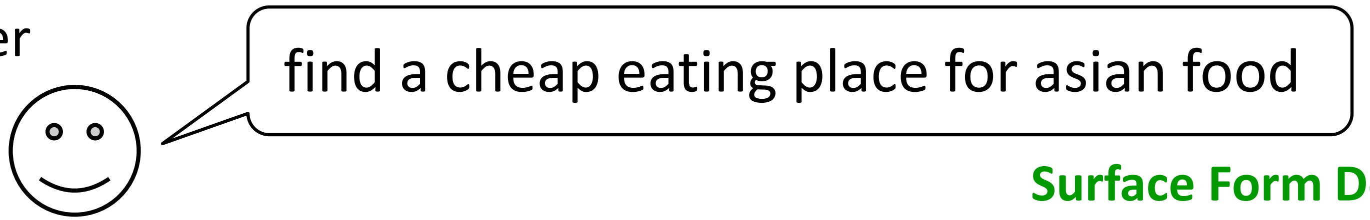
Semantic Decoding

Predicted intent: navigation

Intent Prediction

Thesis Contributions

User



Surface Form Derivation

Ontology Induction

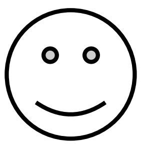
Semantic Decoding

Intent Prediction

Structure Learning

Thesis Contributions

User



find a cheap eating place for asian food

- ✓ Ontology Induction
- ✓ Structure Learning
- ✓ Surface Form Derivation

Knowledge Acquisition

- ✓ Semantic Decoding
- ✓ Intent Prediction

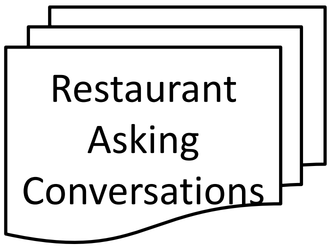
SLU Modeling

Knowledge Acquisition

1) Given unlabelled conversations, how can a system automatically induce and organize domain-specific concepts?

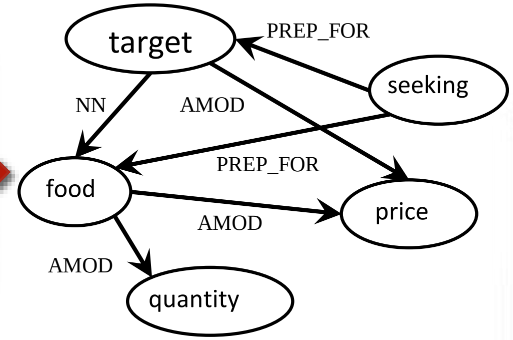
Knowledge Acquisition

- ✓ **Ontology Induction**
- ✓ **Structure Learning**
- ✓ **Surface Form Derivation**



Unlabelled Collection

Knowledge Acquisition



Organized Domain Knowledge

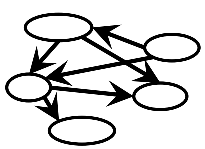
SLU Modeling

2) With the automatically acquired knowledge, how can a system understand utterance semantics and user intents?

SLU Modeling

- ✓ **Semantic Decoding**
- ✓ **Intent Prediction**

"can i have a cheap restaurant"



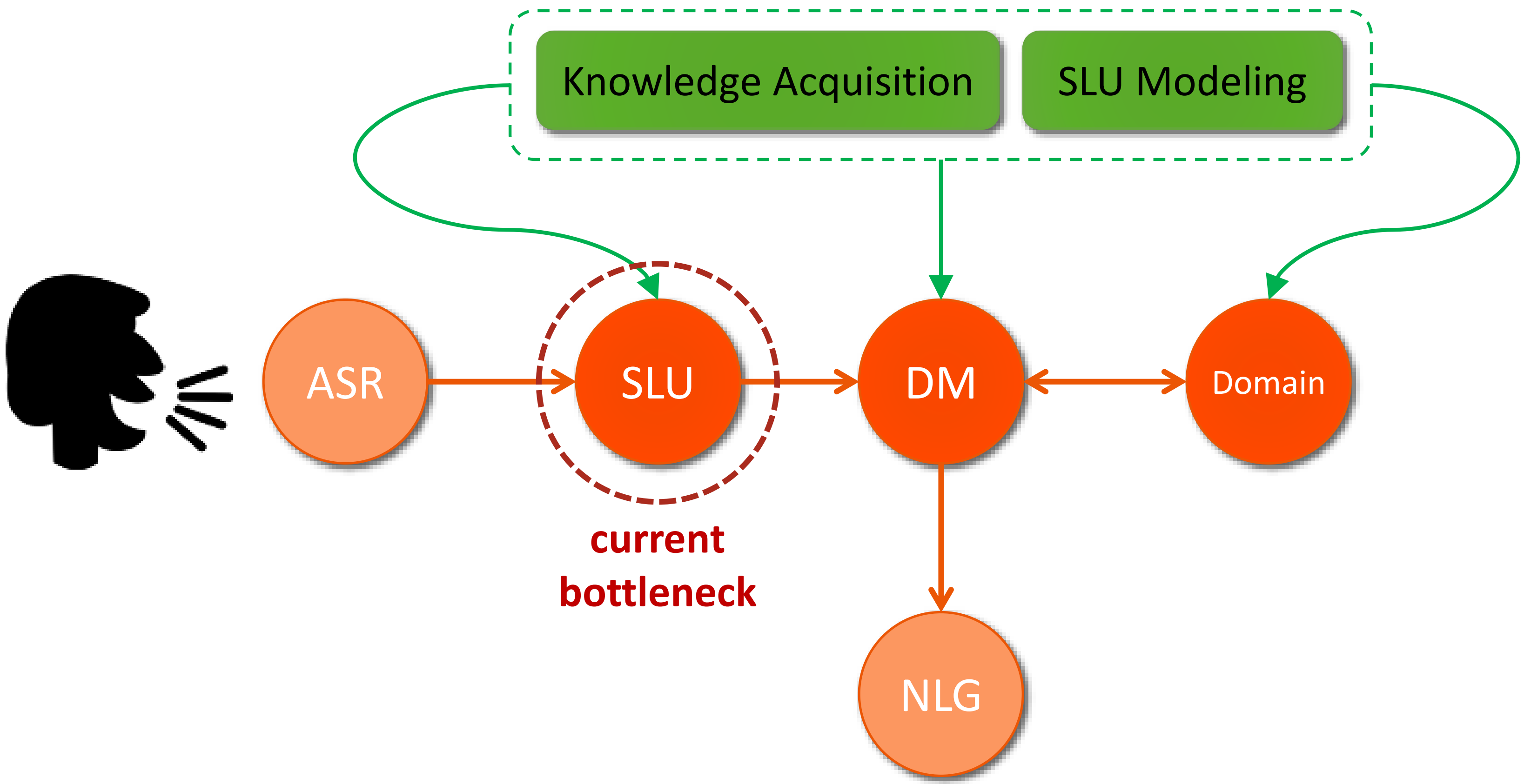
Organized Domain Knowledge

SLU Modeling

SLU Component

price=*"cheap"*
target=*"restaurant"*
intent=*navigation*

SDS Architecture – Contributions



Outline

Introduction

Ontology Induction [ASRU'13, SLT'14a]

Structure Learning [NAACL-HLT'15]

Surface Form Derivation [SLT'14b]

Knowledge Acquisition

Semantic Decoding [ACL-IJCNLP'15]

Intent Prediction [SLT'14c, ICMI'15]

SLU Modeling

SLU in Human-Human Conversations [ASRU'15]

Conclusions & Future Work

Outline

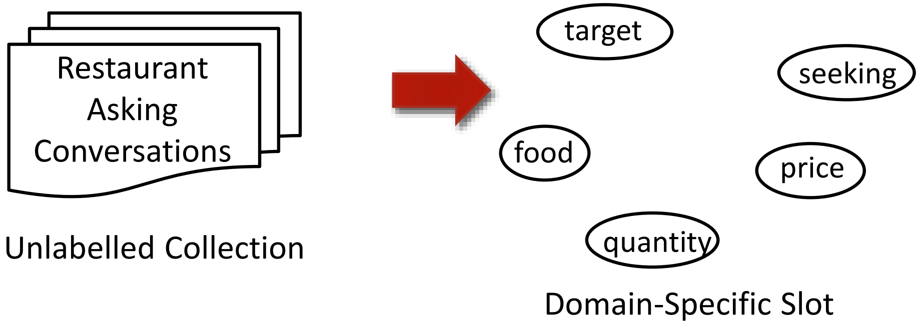
- Introduction
- **Ontology Induction [ASRU'13, SLT'14a]**
- Structure Learning [NAACL-HLT'15]
- Surface Form Derivation [SLT'14b]
- Semantic Decoding [ACL-IJCNLP'15]
- Intent Prediction [SLT'14c, ICMI'15]
- SLU in Human-Human Conversations [ASRU'15]
- Conclusions & Future Work

Ontology Induction

Best Student Paper Award
[ASRU'13, SLT'14a]

Input: Unlabelled user utterances

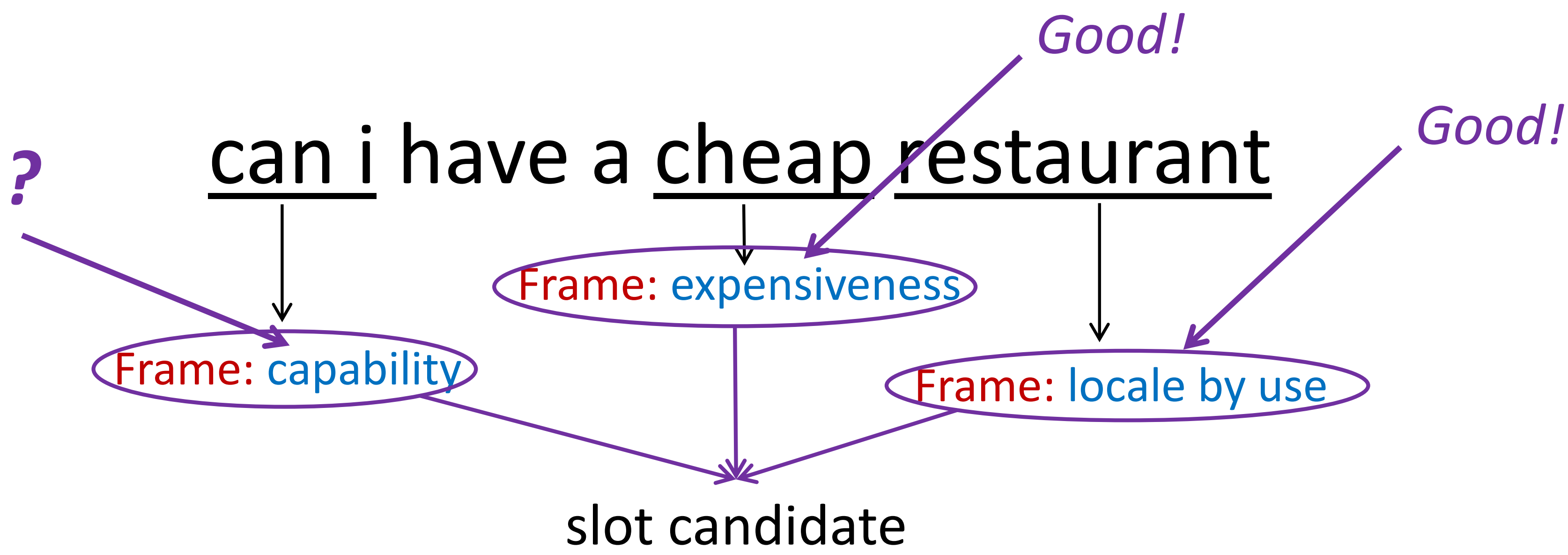
Output: Slots that are useful for a domain-specific SDS



Idea: select a subset of FrameNet-based slots for domain-specific SDS

Chen et al., "Unsupervised Induction and Filling of Semantic Slots for Spoken Dialogue Systems Using Frame-Semantic Parsing," in *Proc. of ASRU,* 2013. **(Best Student Paper Award)**
Chen et al., "Leveraging Frame Semantics and Distributional Semantics for Unsupervised Semantic Slot Induction in Spoken Dialogue Systems," in *Proc. of SLT,* 2014.

Step 1: Frame-Semantic Parsing (Das et al., 2014)



Task: differentiate domain-specific frames from generic frames for SDSs

Das et al., " Frame-semantic parsing," in Proc. of Computational Linguistics, 2014.

Step 2: Slot Ranking Model

Compute an importance score of a slot candidate s by

$$w(s) = (1-\alpha)\log f(s) + \alpha \cdot \log h(s)$$

slot frequency in the domain-specific conversation

slots with higher frequency → more important

semantic coherence of slot fillers

domain-specific concepts → fewer topics

measured by cosine similarity between their word embeddings

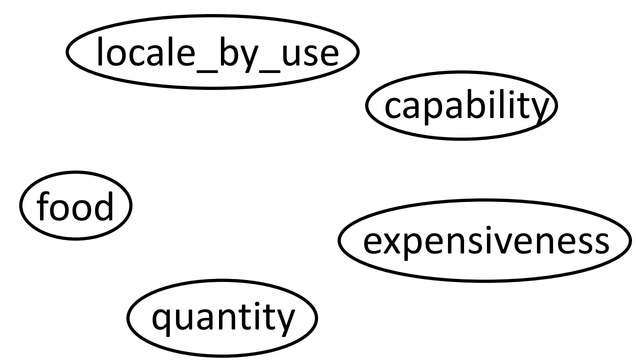
Step 3: Slot Selection

Rank all slot candidates by their importance scores

$$w(s) = (1-\alpha)\log \underbrace{f(s)}_{\text{frequency}} + \alpha \cdot \log \underbrace{h(s)}_{\text{semantic coherence}}$$

Output slot candidates with higher scores based on a threshold

locale_by_use	0.89
capability	0.68
food	0.64
expensiveness	0.49
quantity	0.30
seeking	0.22
:	

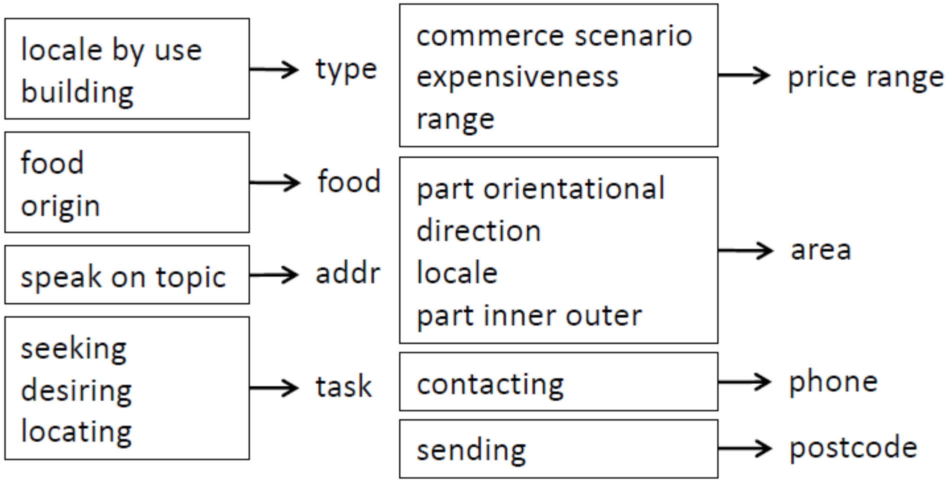


Experiments of Ontology Induction

Dataset

- Cambridge University SLU corpus [Henderson, 2012]
  - Restaurant recommendation (WER = 37%)
    - 2,166 dialogues
    - 15,453 utterances
    - dialogue slot:

addr, area, food, name, phone, postcode, price range, task, type



The mapping table between induced and reference slots

Henderson et al., "Discriminative spoken language understanding using word confusion networks," in *Proc. of SLT*, 2012.

Experiments of Ontology Induction
Experiment: Slot Induction

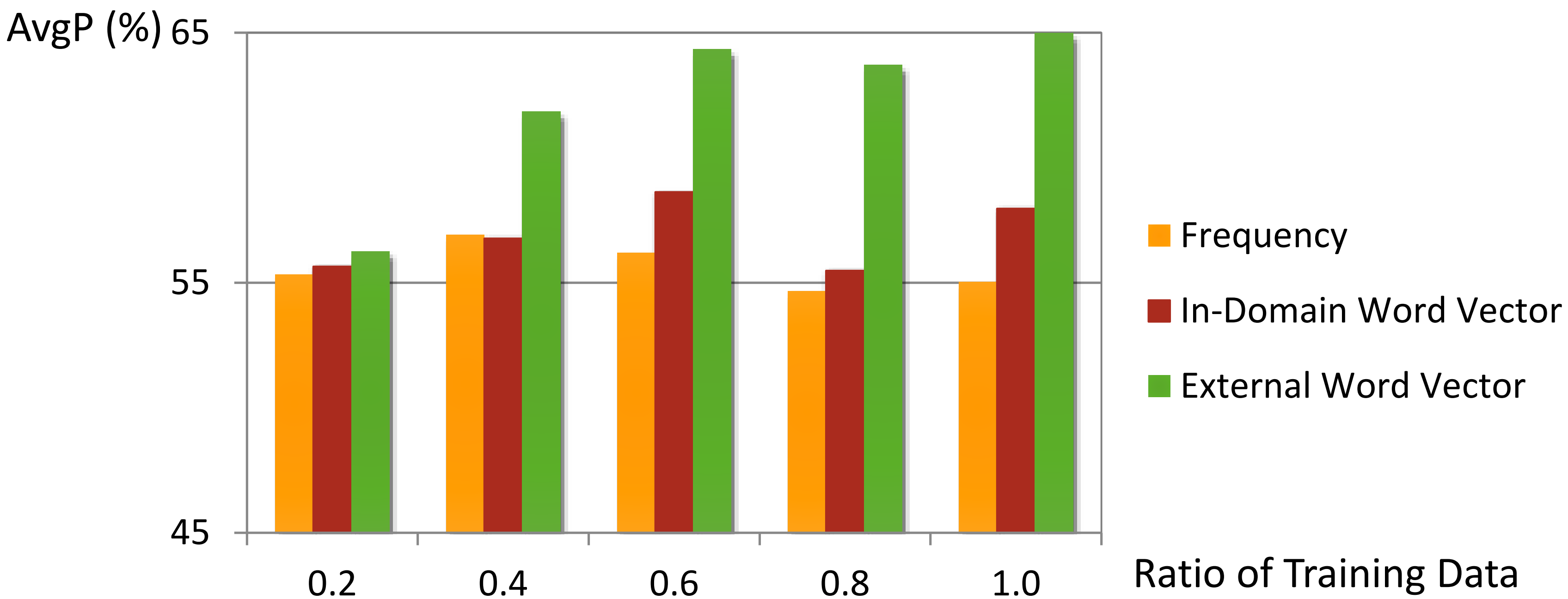
- Metric: Average Precision (AP) and Area Under the Precision-Recall Curve (AUC) of the slot ranking model to measure quality of induced slots via the mapping table

Approach	Word Embedding	ASR		Transcripts	
		AP (%)	AUC (%)	AP (%)	AUC (%)
Baseline: MLE		58.2	56.2	55.0	53.5
Proposed: + Coherence	In-Domain Word Vec.	67.0	65.8	58.0	56.5
	External Word Vec.	**74.5** **(+39.9%)**	**73.5** **(+44.1%)**	**65.0** **(+18.1%)**	**64.2** **(+19.9%)**

Semantic relations help decide domain-specific knowledge.

Experiments of Ontology Induction
Sensitivity to Amount of Training Data

- Different amount of training transcripts for ontology induction



Most approaches are not sensitive to training data size due to <u>single-domain dialogues</u>.

The external word vectors trained on larger data perform better.

Outline

Introduction

Ontology Induction [ASRU'13, SLT'14a]

Structure Learning [NAACL-HLT'15]

Surface Form Derivation [SLT'14b]

Semantic Decoding [ACL-IJCNLP'15]

Intent Prediction [SLT'14c, ICMI'15]

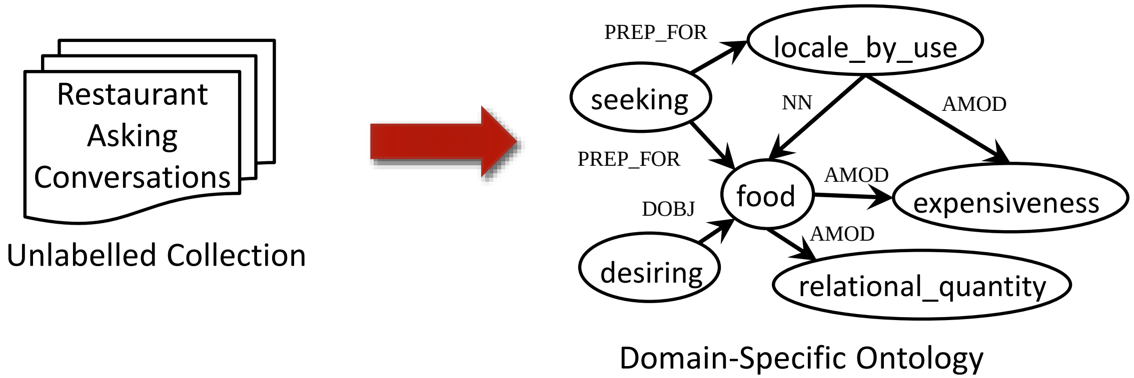
SLU in Human-Human Conversations [ASRU'15]

Conclusions & Future Work

Structure Learning [NAACL-HLT'15]

Input: Unlabelled user utterances

Output: Slots with relations

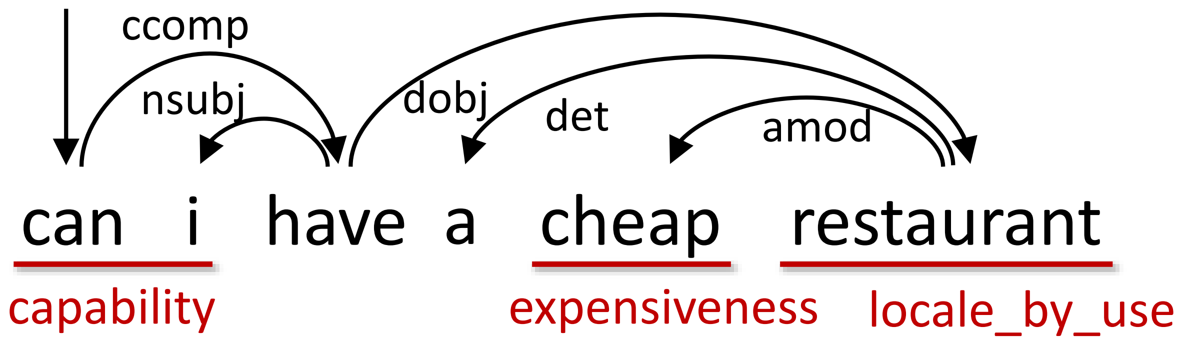


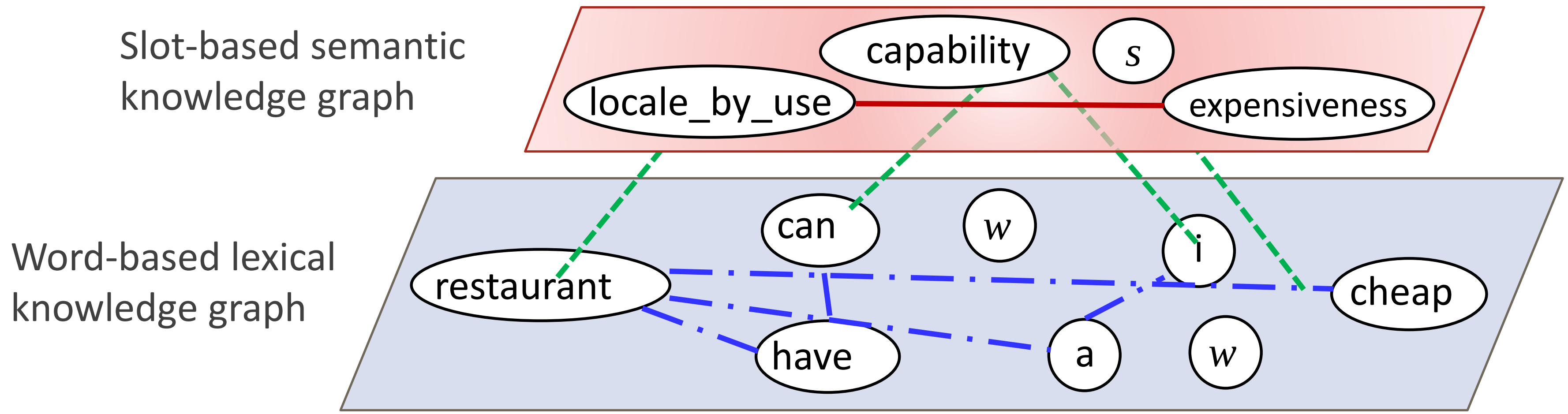
Idea: construct a knowledge graph and then compute slot importance based on relations

Chen et al., "Jointly Modeling Inter-Slot Relations by Random Walk on Knowledge Graphs for Unsupervised Spoken Language Understanding," *in Proc. of NAACL-HLT,* 2015.

Step 1: Knowledge Graph Construction

Syntactic dependency parsing on utterances

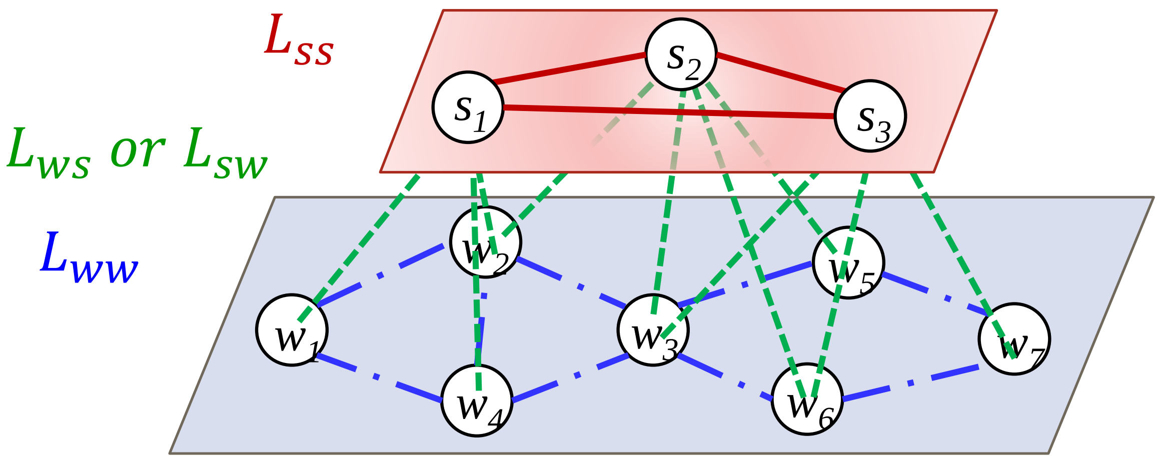


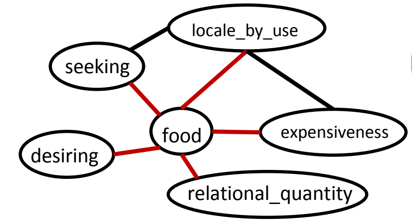


Step 2: Edge Weight Measurement

Compute edge weights to represent relation importance

- Slot-to-slot relation L_{ss}: similarity between slot embeddings
- Word-to-slot relation L_{ws} or L_{sw}: frequency of the slot-word pair
- Word-to-word relation L_{ww}: similarity between word embeddings





Step 2: Slot Importance by Random Walk

Assumption: the slots with more dependencies to more important slots should be more important

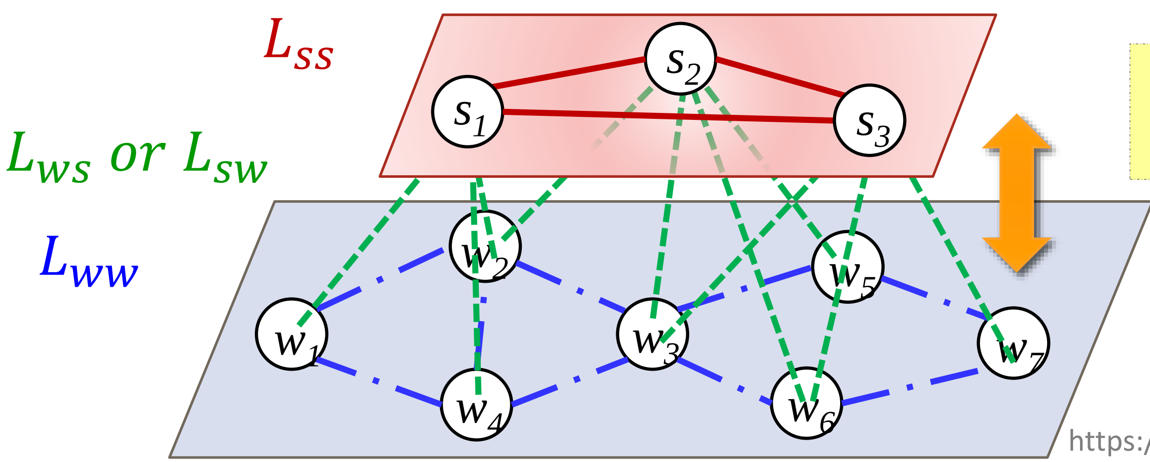
The random walk algorithm computes importance for each slot

slot importance

$$\begin{cases} \boxed{r_s^{(t+1)}} = (1-\alpha)\boxed{r_s^{(0)}} + \alpha \boxed{L_{ss} L_{sw} r_w^{(t)}} \\ r_w^{(t+1)} = (1-\alpha) r_w^{(0)} + \alpha L_{ww} L_{ws} r_s^{(t)} \end{cases}$$

scores propagated from word-layer then propagated within slot-layer

original frequency score



Converged scores can represent the importance.

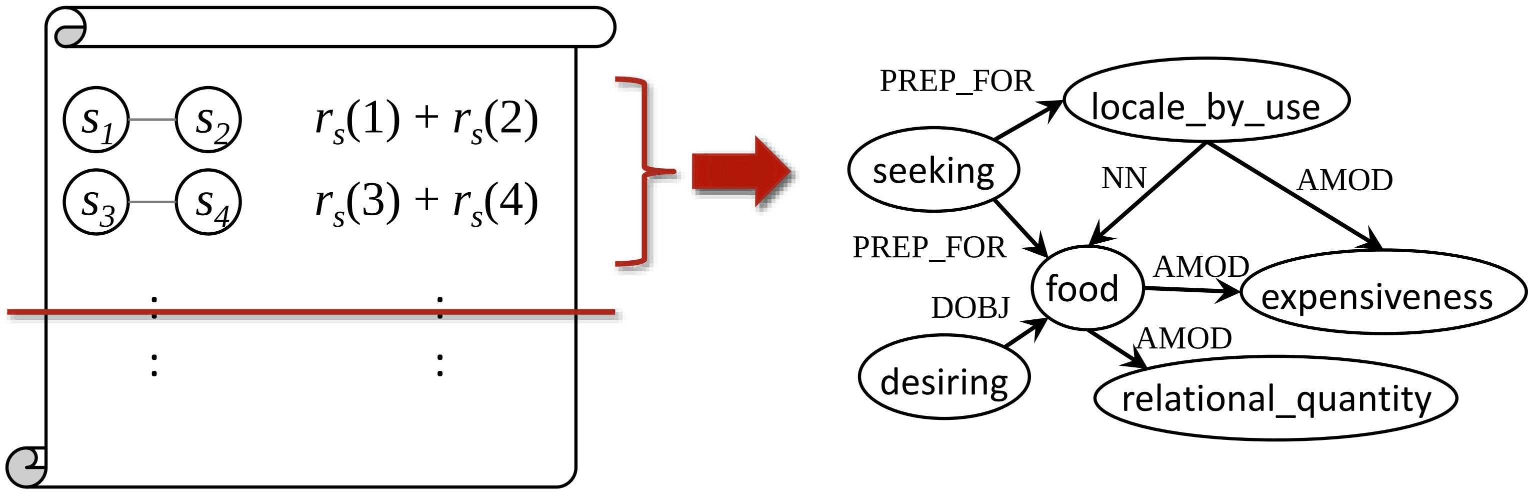
https://github.com/yvchen/MRRW

Step 3: Identify Domain Slots w/ Relations

The converged slot importance suggests whether the slot is important *(Experiment 1)*

Rank slot pairs by summing up their converged slot importance

Select slot pairs with higher scores according to a threshold *(Experiment 2)*



Experiments for Structure Learning

Experiment 1: Quality of Slot Importance

Dataset: Cambridge University SLU Corpus

Approach	ASR		Transcripts	
	AP (%)	AUC (%)	AP (%)	AUC (%)
Baseline: MLE	56.7	54.7	53.0	50.8
Proposed: Random Walk via Dependencies	**71.5** (+26.1%)	**70.8** (+29.4%)	**76.4** (+44.2%)	**76.0** (+49.6%)

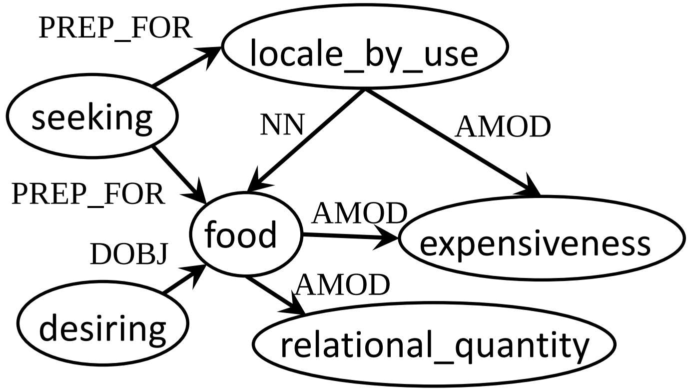
Dependency relations help decide domain-specific knowledge.

Experiments for Structure Learning

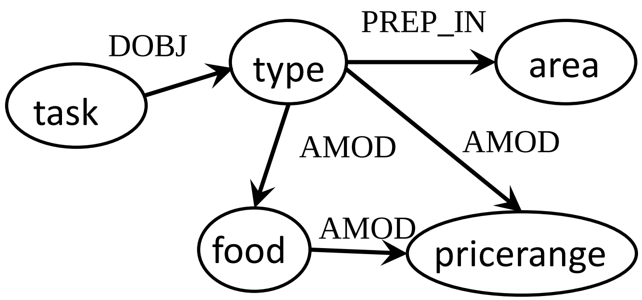
Experiment 2: Relation Discovery Analysis

Discover inter-slot relations connecting important slot pairs

Rank	Relation
1	⟨locale_by_use, NN, food⟩
2	⟨food, AMOD, expensiveness⟩
3	⟨locale_by_use, AMOD, expensiveness⟩
4	⟨seeking, PREP_FOR, food⟩
5	⟨food, AMOD, relational_quantity⟩
6	⟨desiring, DOBJ, food⟩
7	⟨seeking, PREP_FOR, locale_by_use⟩
8	⟨food, DET, quantity⟩



The reference ontology with the most frequent syntactic dependencies

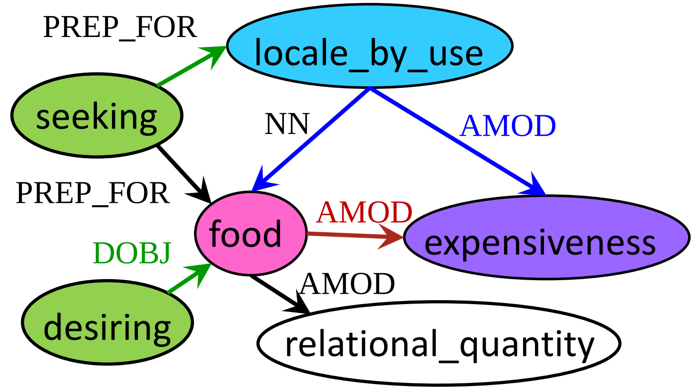


Experiments for Structure Learning

Experiment 2: Relation Discovery Analysis

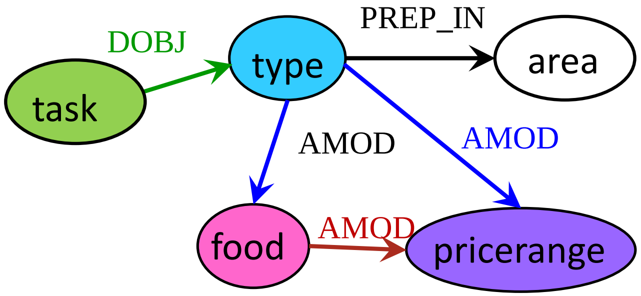
Discover inter-slot relations connecting important slot pairs

Rank	Relation
1	⟨locale_by_use, NN, food⟩
2	⟨food, AMOD, expensiveness⟩
3	⟨locale_by_use, AMOD, expensiveness⟩
4	⟨seeking, PREP_FOR, food⟩
5	⟨food, AMOD, relational_quantity⟩
6	⟨desiring, DOBJ, food⟩
7	⟨seeking, PREP_FOR, locale_by_use⟩
8	⟨food, DET, quantity⟩



The reference ontology with the most frequent syntactic dependencies

The automatically learned domain ontology aligns well with the reference one.



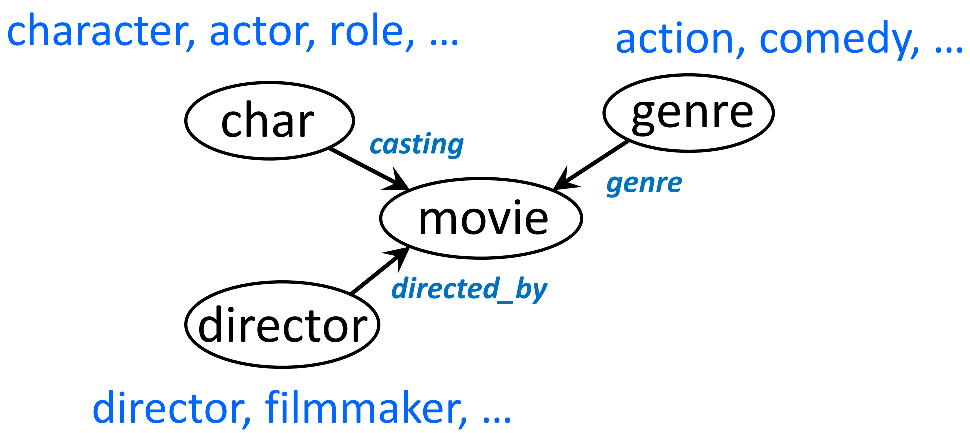
Outline

- Introduction
- Ontology Induction [ASRU'13, SLT'14a]
- Structure Learning [NAACL-HLT'15]
- **Surface Form Derivation [SLT'14b]**
- Semantic Decoding [ACL-IJCNLP'15]
- Intent Prediction [SLT'14c, ICMI'15]
- SLU in Human-Human Conversations [ASRU'15]
- Conclusions & Future Work

Surface Form Derivation [SLT'14b]

Input: a domain-specific organized ontology

Output: surface forms corresponding to entities in the ontology



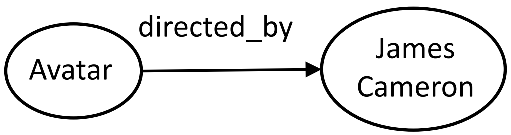
Domain-Specific Ontology

Idea: mine patterns from the web to help understanding

<u>Chen</u> et al., "Deriving Local Relational Surface Forms from Dependency-Based Entity Embeddings for Unsupervised Spoken Language Understanding," in *Proc. of SLT,* 2014.

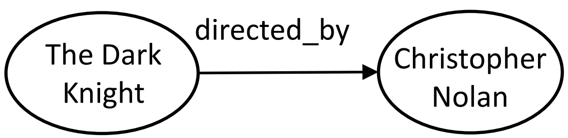
Step 1: Mining Query Snippets on Web

Query snippets including entity pairs connected with specific relations in KG



Avatar is a 2009 American epic science fiction film directed by James Cameron.

directed_by

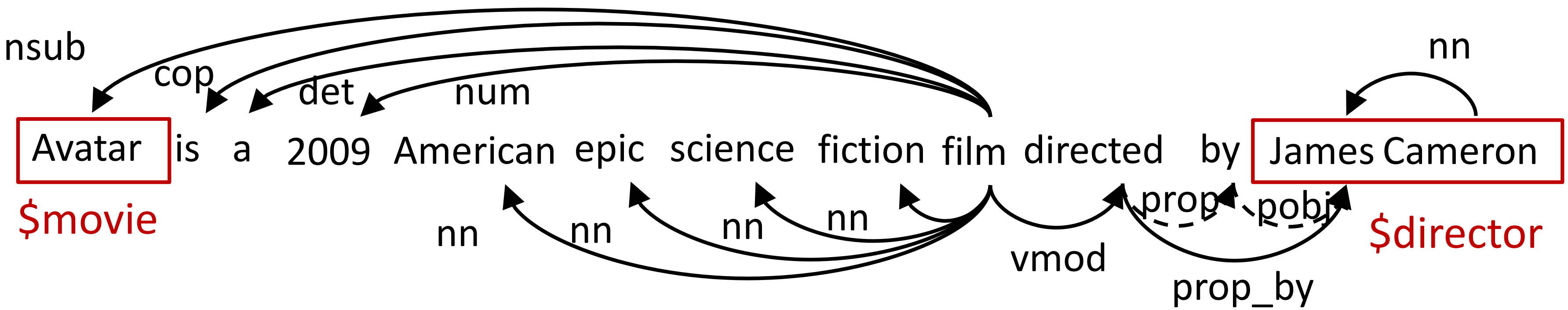


The Dark Knight is a 2008 superhero film directed and produced by Christopher Nolan.

directed_by

Step 2: Training Entity Embeddings

Dependency parsing for training dependency-based embeddings



$movie = [0.8 … 0.24]
is = [0.3 … 0.21]
film = [0.12 … 0.7]
:
:

Levy and Goldberg, " Dependency-Based Word Embeddings," in *Proc. of ACL*, 2014.

Step 3: Deriving Surface Forms

Entity Surface Forms

- learn the surface forms corresponding to entities
- most similar word vectors for each entity embedding

$char: “character”, “role”, “who”
$director: “director”, “filmmaker”
$genre: “action”, “fiction”

→ with similar contexts

Entity Contexts

- learn the important contexts of entities
- most similar context vectors for each entity embedding

$char: “played”
$director: “directed”

→ frequently occurring together

Experiments of Surface Form Derivation

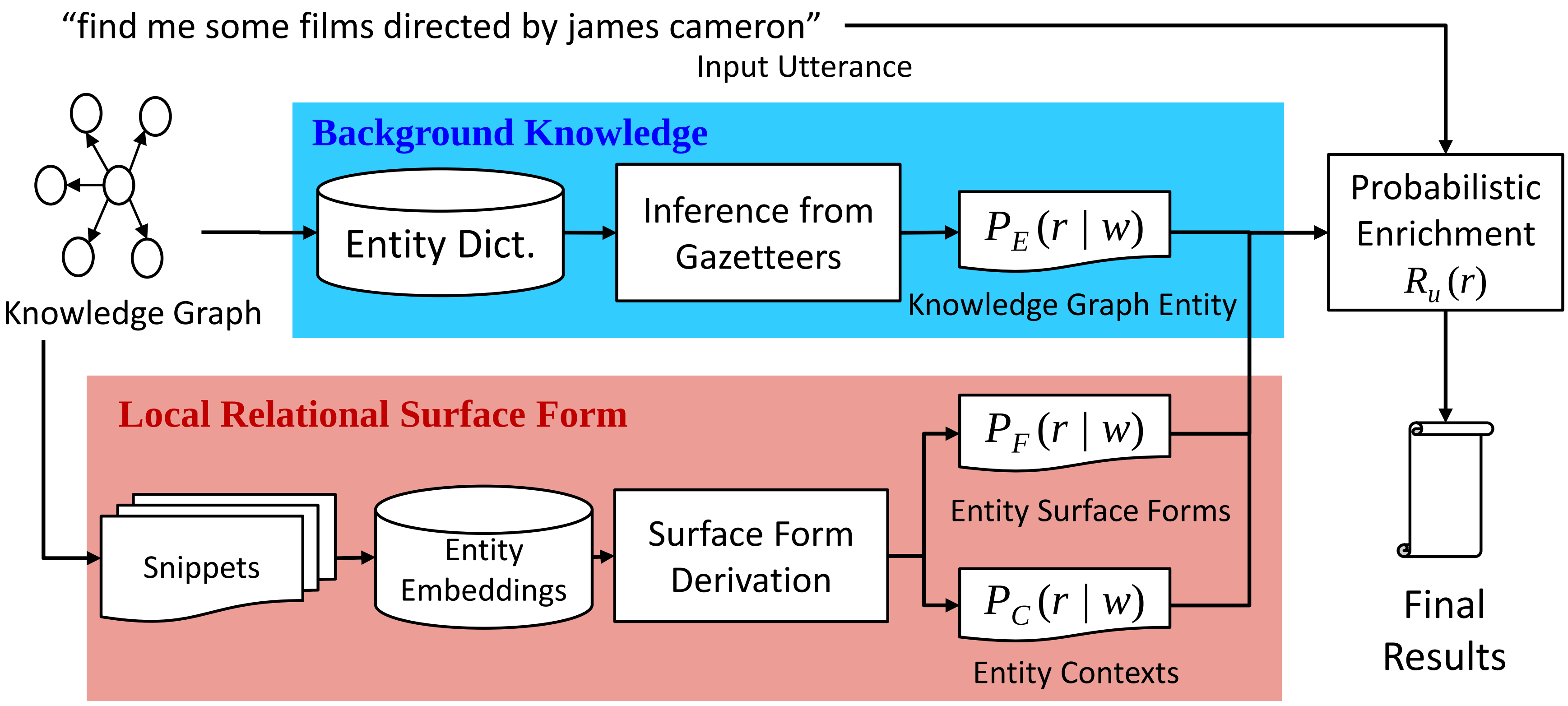
Knowledge Base: Freebase

- 670K entities; 78 entity types (movie names, actors, etc)

Entity Tag	Derived Word
$character	character, role, who, girl, she, he, officier
$director	director, dir, filmmaker
$genre	comedy, drama, fantasy, cartoon, horror, sci
$language	language, spanish, english, german
$producer	producer, filmmaker, screenwriter

The web-derived surface forms provide useful knowledge for better understanding.

Integrated with Background Knowledge



Hakkani-Tür et al., "Probabilistic enrichment of knowledge graph entities for relation detection in conversational understanding," in *Proc. of Interspeech,* 2014.

Experiments of Surface Form Derivation

Relation Detection Data (NL-SPARQL): a dialog system challenge set for converting natural language to structured queries (Hakkani-Tür et al., 2014)

- Crowd-sourced utterances (3,338 for self-training, 1,084 for testing)

Metric: micro F-measure (%)

Approach	Micro F-Measure (%)
Baseline: Gazetteer	38.9
Gazetteer + Entity Surface Form + Entity Context	**43.3** **(+11.4%)**

The web-derived knowledge can benefit SLU performance.

Hakkani-Tür et al., "Probabilistic enrichment of knowledge graph entities for relation detection in conversational understanding," in *Proc. of Interspeech,* 2014.

Outline

- Introduction
- Ontology Induction [ASRU'13, SLT'14a]
- Structure Learning [NAACL-HLT'15]
- Surface Form Derivation [SLT'14b]
- Semantic Decoding [ACL-IJCNLP'15]
- Intent Prediction [SLT'14c, ICMI'15]
- SLU in Human-Human Conversations [ASRU'15]
- Conclusions & Future Work

Knowledge Acquisition

SLU Modeling

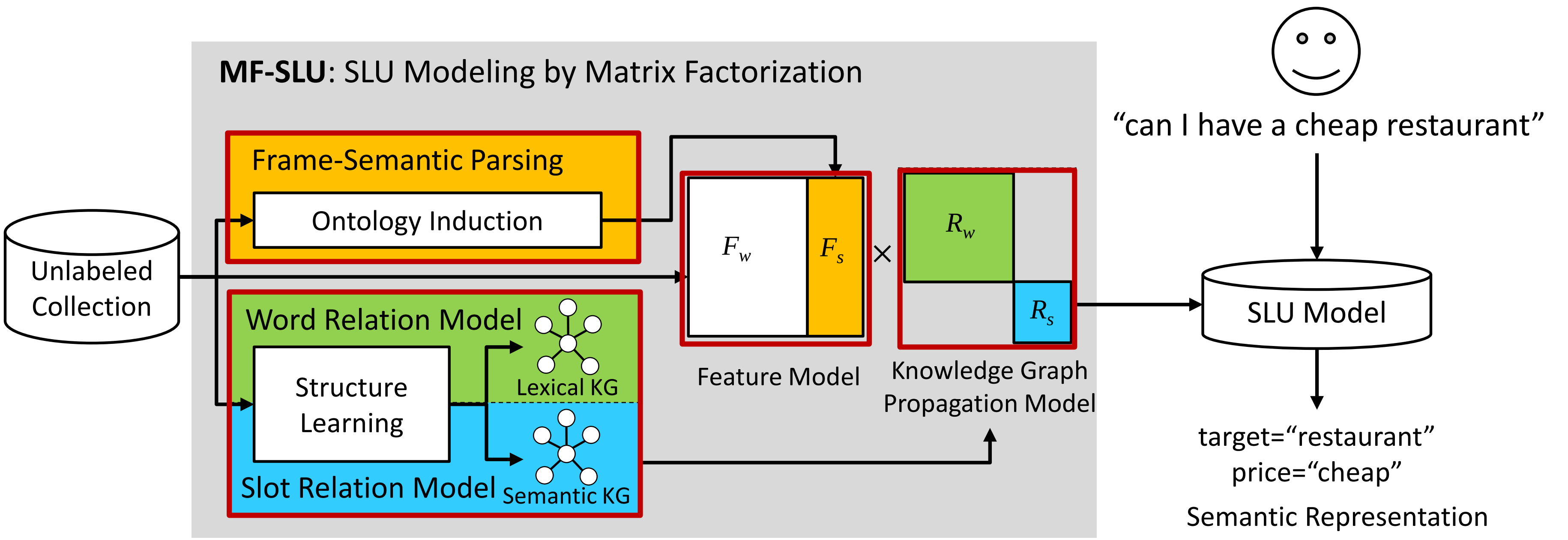
Outline

- Introduction
- Ontology Induction [ASRU'13, SLT'14a]
- Structure Learning [NAACL-HLT'15]
- Surface Form Derivation [SLT'14b]
- **Semantic Decoding [ACL-IJCNLP'15]**
- Intent Prediction [SLT'14c, ICMI'15]
- SLU in Human-Human Conversations [ASRU'15]
- Conclusions & Future Work

Semantic Decoding [ACL-IJCNLP'15]

Input: user utterances, automatically acquired knowledge

Output: the semantic concepts included in each individual utterance

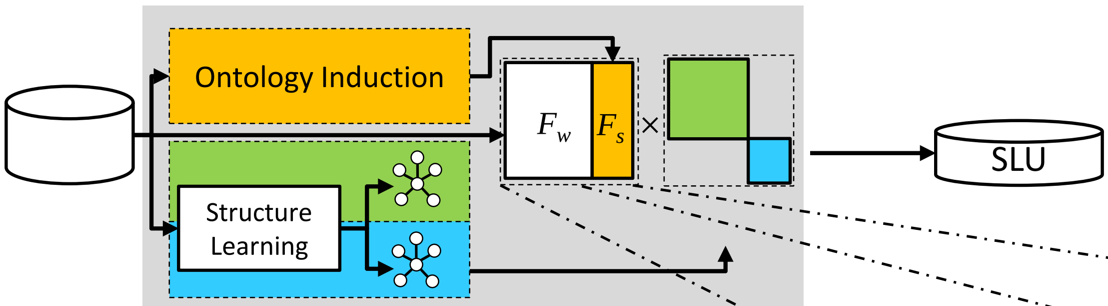


Idea: utilize the acquired knowledge to decode utterance semantics **(fully unsupervised)**

Chen et al., "Matrix Factorization with Knowledge Graph Propagation for Unsupervised Spoken Language Understanding," in *Proc. of ACL-IJCNLP*, 2015.

Matrix Factorization SLU (MF-SLU)

Feature Model



Utterance 1

i would like a cheap restaurant

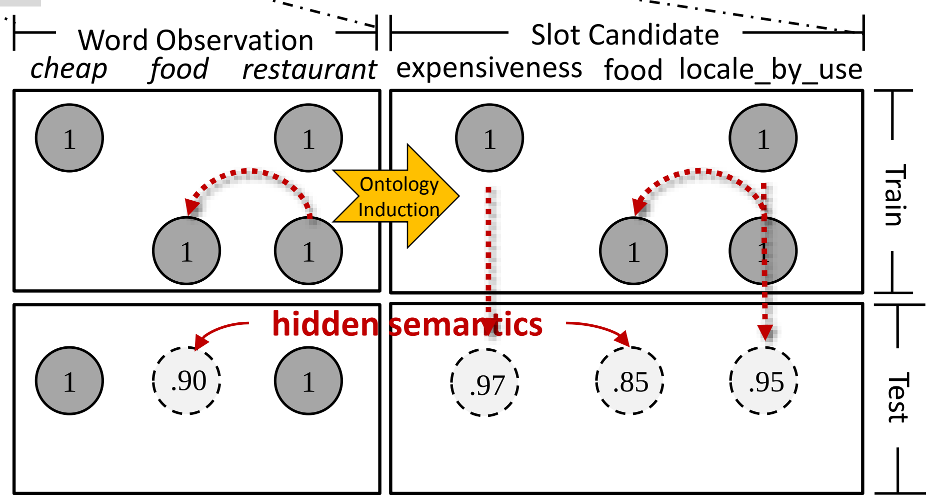
Utterance 2

find a restaurant with chinese food

Test Utterance

show me a list of cheap restaurants

⋮

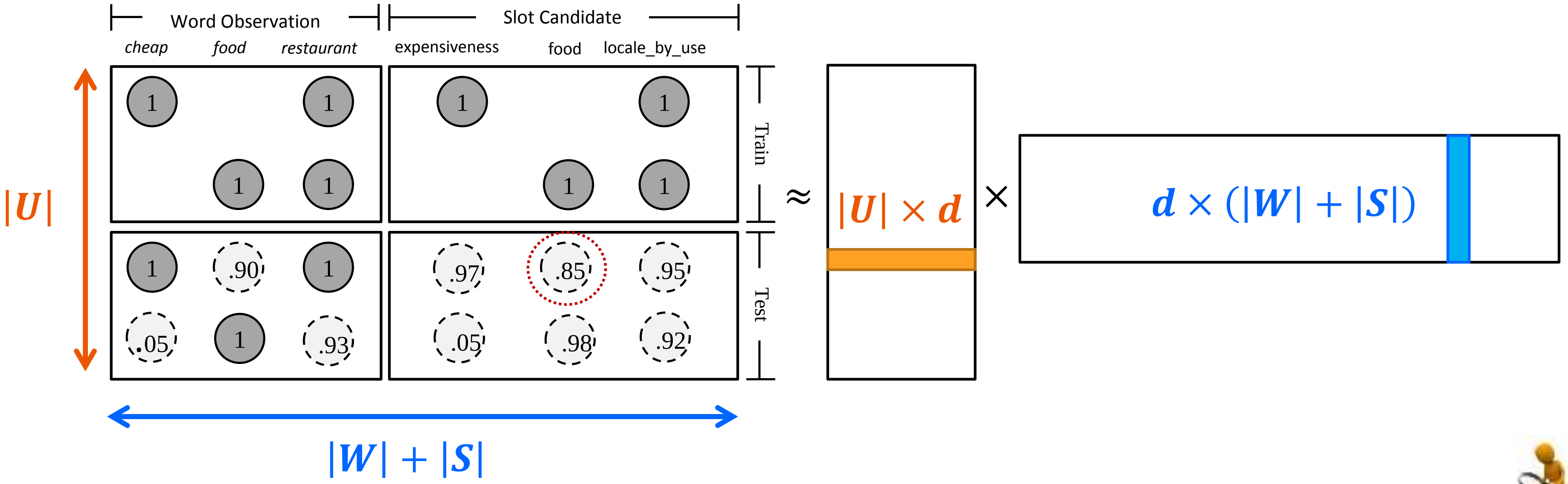


MF completes a partially-missing matrix based on a low-rank latent semantics assumption.

Matrix Factorization (Rendle et al., 2009)

The decomposed matrices represent latent semantics for utterances and words/slots respectively

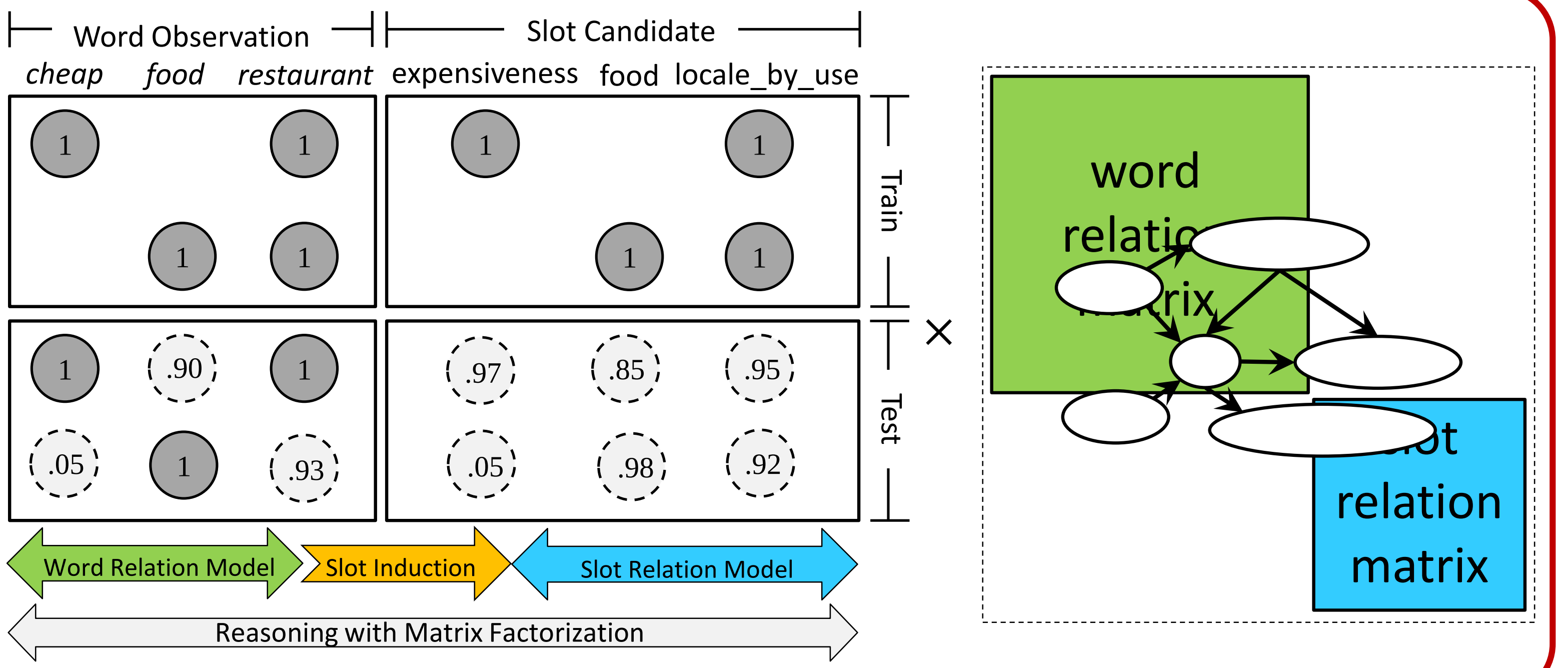
The product of two matrices fills the probability of hidden semantics



Rendle et al., "BPR: Bayesian Personalized Ranking from Implicit Feedback," in *Proc. of UAI*, 2009.

Matrix Factorization SLU (MF-SLU)

Feature Model + Knowledge Graph Propagation Model



Structure information is integrated to make the self-training data more reliable before MF.

Experiments of Semantic Decoding
Experiment 1: Quality of Semantics Estimation

Dataset: Cambridge University SLU Corpus

Metric: MAP of all estimated slot probabilities for each utterance

Approach		ASR	Transcripts
Baseline: SLU	Support Vector Machine	32.5	36.6
	Multinomial Logistic Regression	34.0	38.8

Experiments of Semantic Decoding

Experiment 1: Quality of Semantics Estimation

Dataset: Cambridge University SLU Corpus

Metric: MAP of all estimated slot probabilities for each utterance

Approach		ASR	Transcripts
Baseline: SLU	Support Vector Machine	32.5	36.6
	Multinomial Logistic Regression	34.0	38.8
Proposed: MF-SLU	Feature Model	37.6*	45.3*
	Feature Model + Knowledge Graph Propagation	**43.5*** (+27.9%)	**53.4*** (+37.6%)

The MF-SLU effectively models implicit information to decode semantics.

The structure information further improves the results.

*: the result is significantly better than the MLR with $p < 0.05$ in t-test

Experiments of Semantic Decoding

Experiment 2: Effectiveness of Relations

Dataset: Cambridge University SLU Corpus

Metric: MAP of all estimated slot probabilities for each utterance

Approach		ASR	Transcripts
Feature Model		37.6	45.3
Feature + Knowledge Graph Propagation	Semantic	41.4*	51.6*
	Dependency	41.6*	49.0*
	All	**43.5*** **(+15.7%)**	**53.4*** **(+17.9%)**

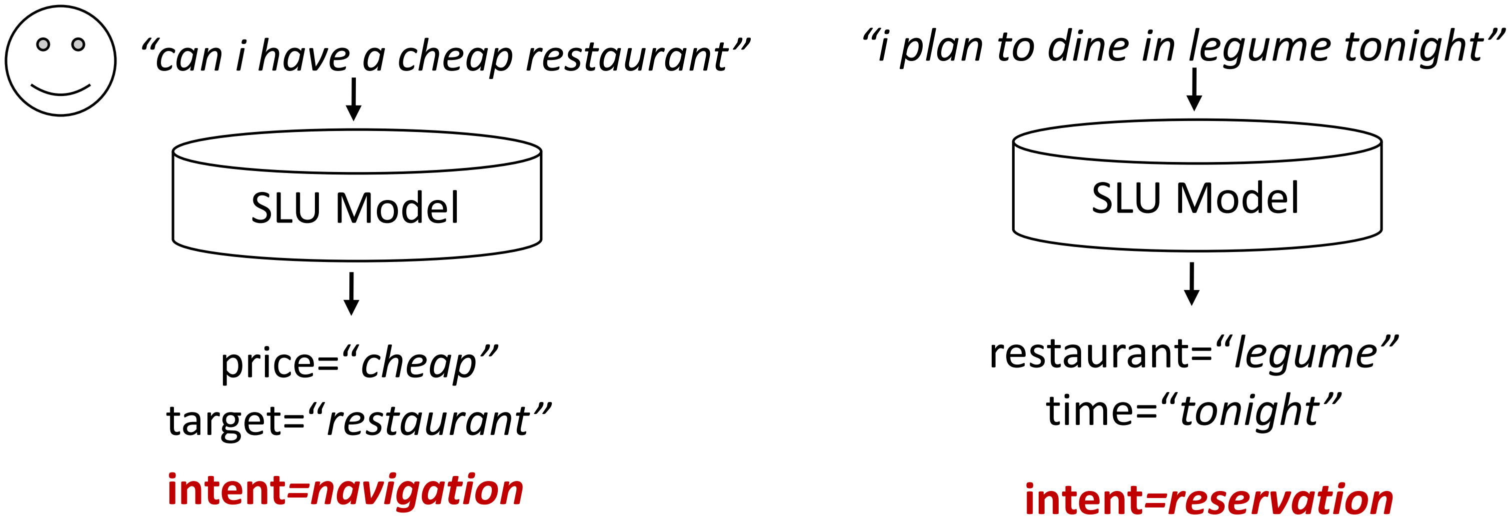
In the integrated structure information, both semantic and dependency relations are useful for understanding.

*: the result is significantly better than the MLR with $p < 0.05$ in t-test

Low- and High-Level Understanding

Semantic concepts for individual utterances do not consider high-level semantics (user intents)

The follow-up behaviors usually correspond to user intents



Outline

- Introduction
- Ontology Induction [ASRU'13, SLT'14a]
- Structure Learning [NAACL-HLT'15]
- Surface Form Derivation [SLT'14b]
- Semantic Decoding [ACL-IJCNLP'15]
- **Intent Prediction [SLT'14c, ICMI'15]**
- SLU in Human-Human Conversations [ASRU'15]
- Conclusions & Future Work

Intent Prediction of Mobile Apps [SLT'14c]

Input: spoken utterances for making requests about launching an app

Output: the apps supporting the required functionality

Intent Identification

- popular domains in Google Play

please dial a phone call to alex

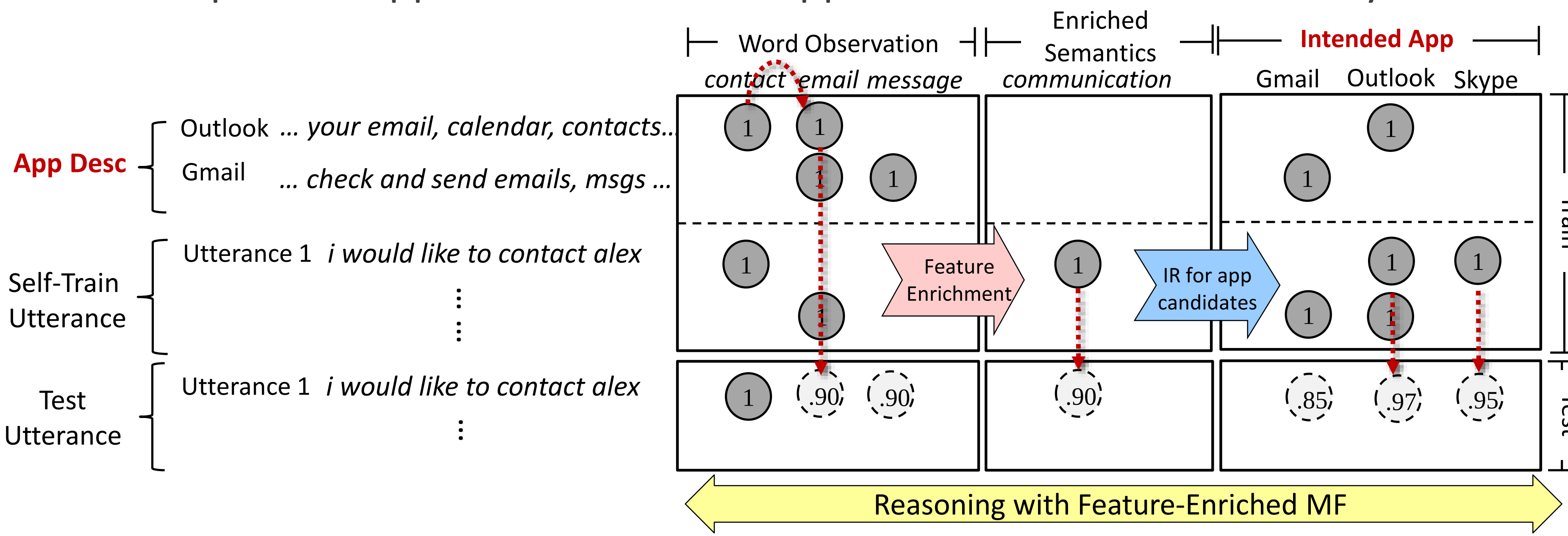
Skype, Hangout, etc.

Chen and Rudnicky, "Dynamically Supporting Unexplored Domains in Conversational Interactions by Enriching Semantics with Neural Word Embeddings," in *Proc. of SLT*, 2014.

Intent Prediction – Single-Turn

Input: single-turn request

Output: the apps that are able to support the required functionality



The *feature-enriched MF-SLU* unifies manually written knowledge and automatically inferred semantics to predict high-level intents.

Intent Prediction – Multi-Turn Interaction [ICMI'15]

Input: multi-turn interaction

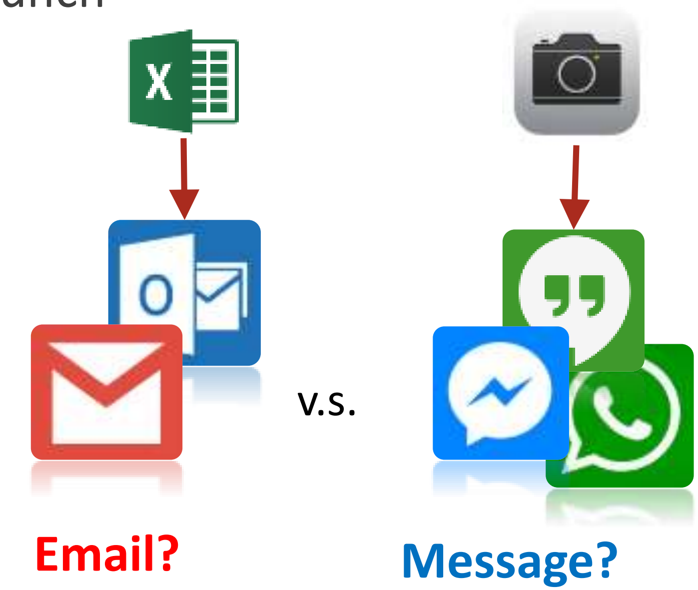
Output: the app the user plans to launch

Challenge: language ambiguity

1) User preference
2) App-level contexts

send to vivian

Communication



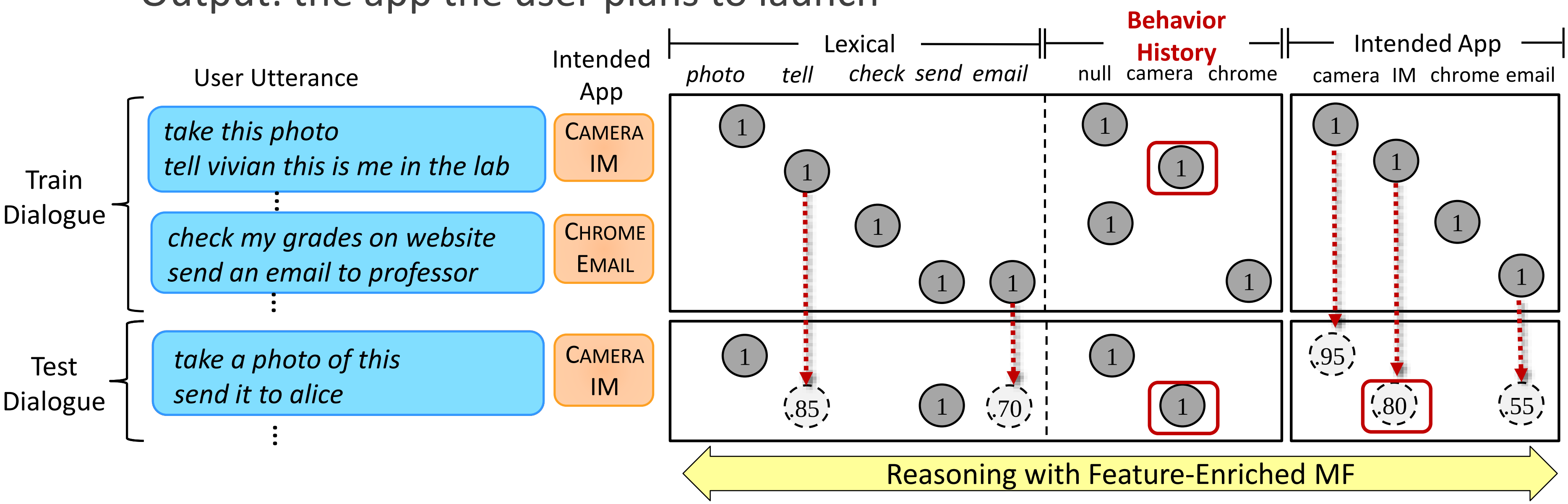
Idea: Behavioral patterns in history can help intent prediction.

Chen et al., "Leveraging Behavioral Patterns of Mobile Applications for Personalized Spoken Language Understanding," in *Proc. of ICMI*, 2015. Data Available at http://AppDialogue.com/.

Intent Prediction – Multi-Turn Interaction [ICMI'15]

Input: multi-turn interaction

Output: the app the user plans to launch



The *feature-enriched MF-SLU* leverages behavioral patterns to model contextual information and user preference for better intent prediction.

Chen et al., "Leveraging Behavioral Patterns of Mobile Applications for Personalized Spoken Language Understanding," in *Proc. of ICMI*, 2015. Data Available at http://AppDialogue.com/.

Experiments for Intent Prediction

Single-Turn Request: Mean Average Precision (MAP)

LM-Based IR Model (unsupervised)

Feature Matrix	ASR		Transcripts	
	LM	MF-SLU	LM	MF-SLU
Word Observation	25.1		26.1	

Multinomial Logistic Regression (supervised)

Multi-Turn Interaction: Mean Average Precision (MAP)

Feature Matrix	ASR		Transcripts	
	MLR	MF-SLU	MLR	MF-SLU
Word Observation	52.1		55.5	

Experiments for Intent Prediction

Single-Turn Request: Mean Average Precision (MAP)

Feature Matrix	ASR		Transcripts	
	LM	MF-SLU	LM	MF-SLU
Word Observation	25.1	**29.2** (+16.2%)	26.1	**30.4** (+16.4%)

Multi-Turn Interaction: Mean Average Precision (MAP)

Feature Matrix	ASR		Transcripts	
	MLR	MF-SLU	MLR	MF-SLU
Word Observation	52.1	**52.7** (+1.2%)	**55.5**	55.4 (-0.2%)

Modeling hidden semantics helps intent prediction especially for noisy data.

Experiments for Intent Prediction

Single-Turn Request: Mean Average Precision (MAP)

Feature Matrix	ASR		Transcripts	
	LM	MF-SLU	LM	MF-SLU
Word Observation	25.1	29.2 (+16.2%)	26.1	30.4 (+16.4%)
Word + Embedding-Based Semantics	**32.0**		**33.3**	
Word + Type-Embedding-Based Semantics	31.5		32.9	

Multi-Turn Interaction: Mean Average Precision (MAP)

Feature Matrix	ASR		Transcripts	
	MLR	MF-SLU	MLR	MF-SLU
Word Observation	52.1	52.7 (+1.2%)	55.5	55.4 (-0.2%)
Word + Behavioral Patterns	**53.9**		**56.6**	

Semantic enrichment provides rich cues to improve performance.

Experiments for Intent Prediction

Single-Turn Request: Mean Average Precision (MAP)

Feature Matrix	ASR		Transcripts	
	LM	MF-SLU	LM	MF-SLU
Word Observation	25.1	29.2 (+16.2%)	26.1	30.4 (+16.4%)
Word + Embedding-Based Semantics	32.0	**34.2 (+6.8%)**	33.3	33.3 (-0.2%)
Word + Type-Embedding-Based Semantics	31.5	32.2 (+2.1%)	32.9	**34.0 (+3.4%)**

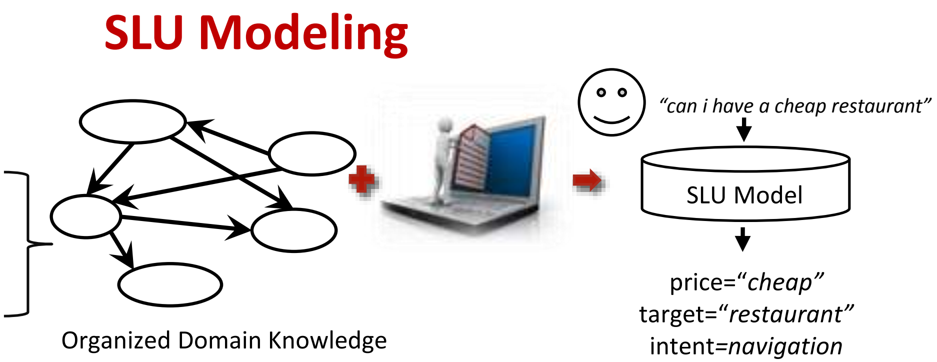
Multi-Turn Interaction: Mean Average Precision (MAP)

Feature Matrix	ASR		Transcripts	
	MLR	MF-SLU	MLR	MF-SLU
Word Observation	52.1	52.7 (+1.2%)	55.5	55.4 (-0.2%)
Word + Behavioral Patterns	53.9	**55.7 (+3.3%)**	56.6	**57.7 (+1.9%)**

Intent prediction can benefit from both hidden information and low-level semantics.

Outline

- Introduction
- Ontology Induction [ASRU'13, SLT'14a]
- Structure Learning [NAACL-HLT'15]
- Surface Form Derivation [SLT'14b]
- Semantic Decoding [ACL-IJCNLP'15]
- Intent Prediction [SLT'14c, ICMI'15]
- SLU in Human-Human Conversations [ASRU'15]
- Conclusions & Future Work



Outline

Introduction

Ontology Induction [ASRU'13, SLT'14a]

Structure Learning [NAACL-HLT'15]

Surface Form Derivation [SLT'14b]

Semantic Decoding [ACL-IJCNLP'15]

Intent Prediction [SLT'14c, ICMI'15]

SLU in Human-Human Conversations [ASRU'15]

Conclusions & Future Work

SLU in Human-Human Dialogues

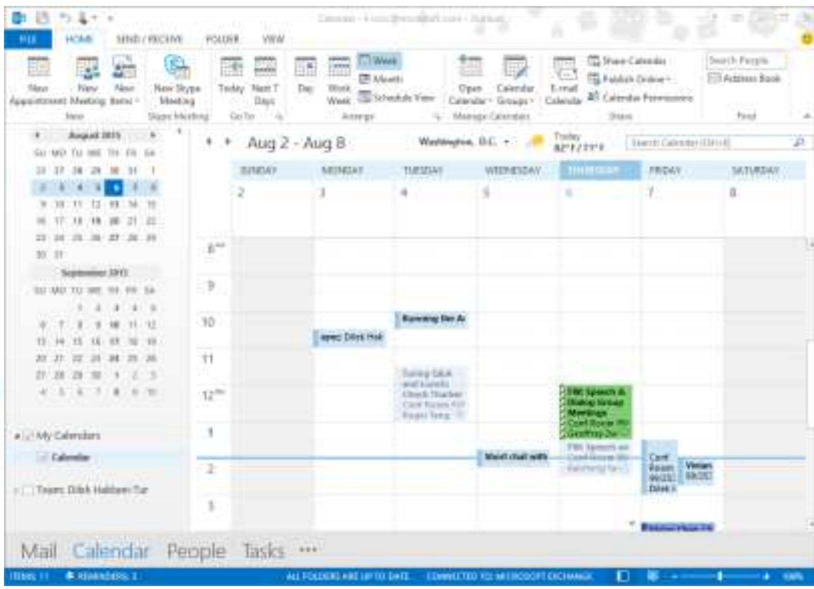
Computing devices have been easily accessible during regular <u>human-human conversations</u>.

The dialogues include discussions for identifying speakers' next actions.

Is it possible to apply the techniques developed for human-machine dialogues to human-human dialogues?

Actionable Item Utterances

Will Vivian come here for the meeting?



Find Calendar Entry

Actionable Item Detection [ASRU'15]

Goal: provide actions an existing system can handle w/o interrupting conversations

Assumption: some actions and associated arguments can be shared across genres

Human-Machine Genre create_calendar_entry

schedule a meeting with John (contact_name) this afternoon (start_time)

Human-Human Genre create_calendar_entry

how about the three of us (contact_name) discuss this later this afternoon (start_time) ?

→ more casual, include conversational terms

Task: multi-class utterance classification

- train on the available human-machine genre
- test on the human-human genre

Adaptation

- model adaptation
- embedding vector adaptation

Chen et al., "Detecting Actionable Items in Meetings by Convolutional Deep Structred Semantic Models," in *Proc. of ASRU*, 2015.

Action Item Detection

Idea: human-machine interactions may help detect actionable items in human-human conversations

Query

12 30 on Monday the 2nd of July schedule meeting to discuss taxes with Bill, Judy and Rick.
12 PM lunch with Jessie
12:00 meeting with cleaning staff to discuss progress
2 hour business presentation with Rick Sandy on March 8 at 7am

A read receipt should be sent in Karen's email.
Activate meeting delay by 15 minutes. Inform the participants.
Add more message saying "Thanks".

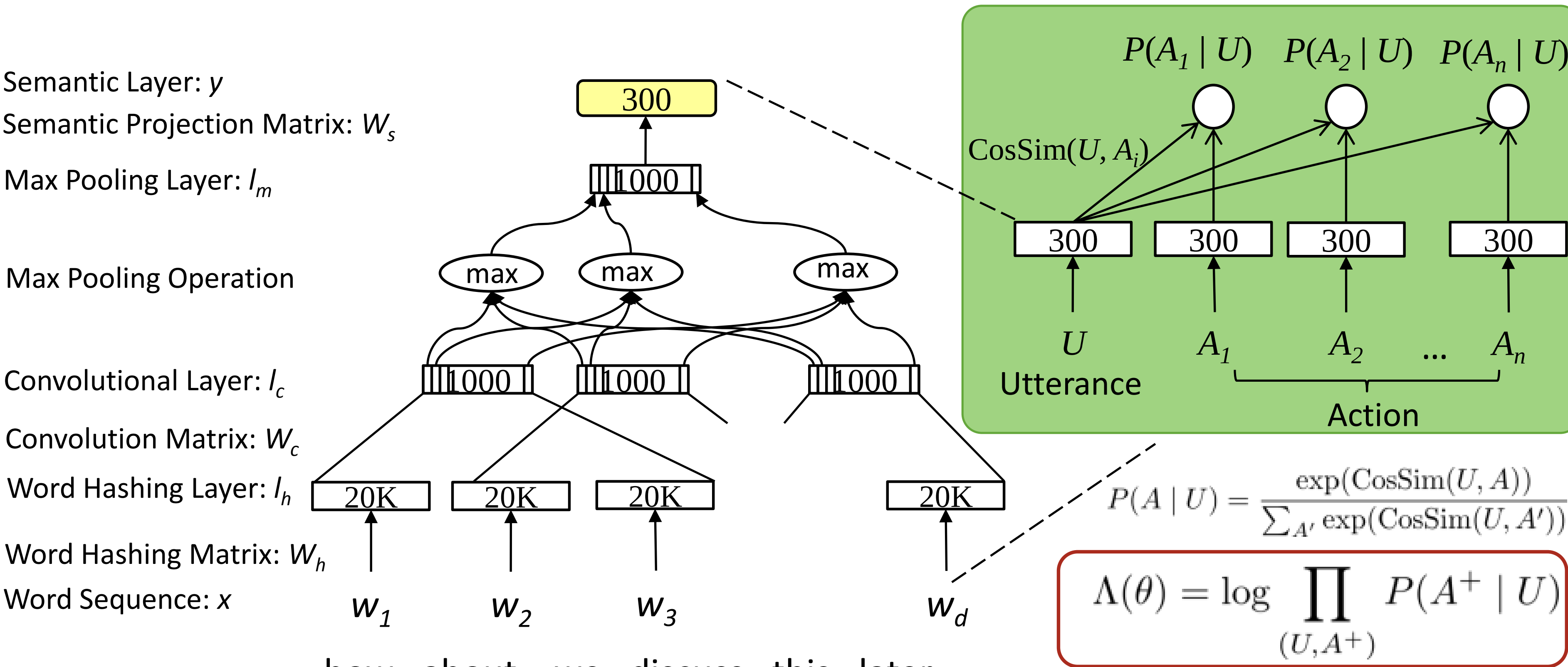
Doc

create_calendar_entry

send_email

Convolutional deep structured semantic models for IR may be useful for this task.

Convolutional Deep Structured Semantic Models (CDSSM) (Huang et al., 2013; Shen et al., 2014)



maximizes the likelihood of associated actions given utterances

Huang et al., "Learning deep structured semantic models for web search using clickthrough data," in *Proc. of CIKM*, 2013.
Shen et al., "Learning semantic representations using ´ convolutional neural networks for web search," in *Proc. of WWW*, 2014.

Adaptation

Issue: mismatch between a source genre and a target genre

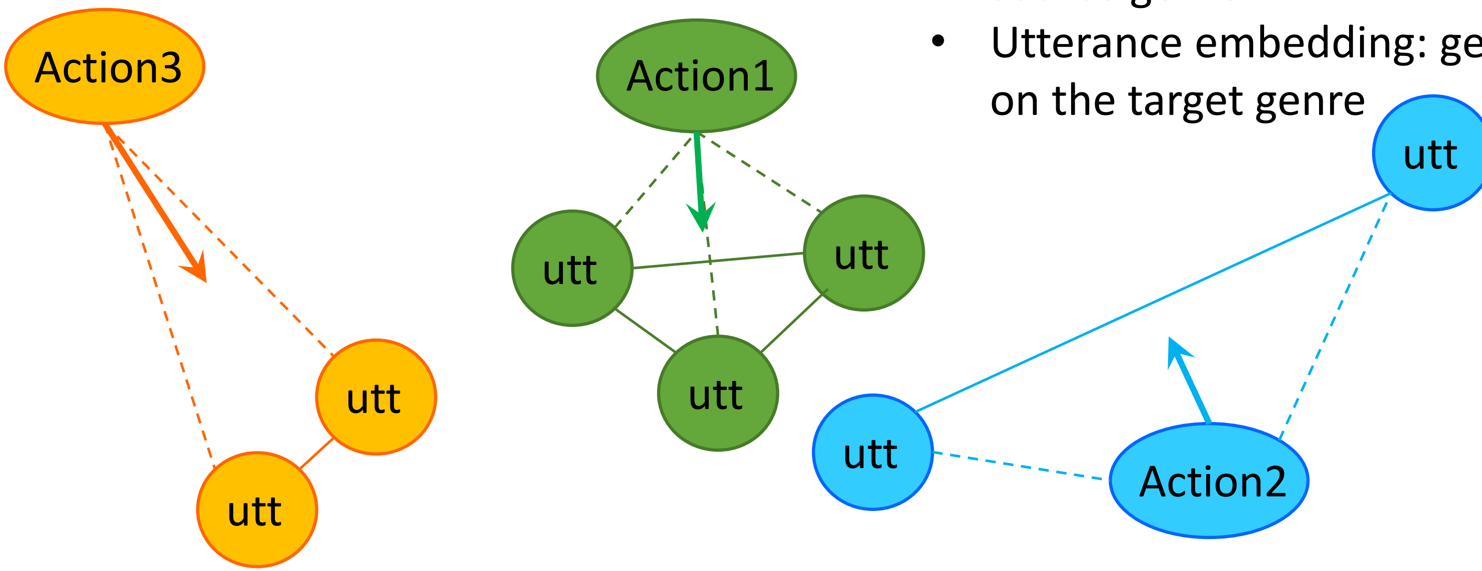
Solution:

- Adapting CDSSM
  - Continually train the CDSSM using the data from the target genre
- Adapting Action Embeddings
  - Moving learned action embeddings close to the observed corresponding utterance embeddings

Adapting Action Embeddings

The mismatch may result in inaccurate action embeddings

- Action embedding: trained on the source genre
- Utterance embedding: generated on the target genre



Idea: moving action embeddings close to the observed utterance embeddings from the target genre

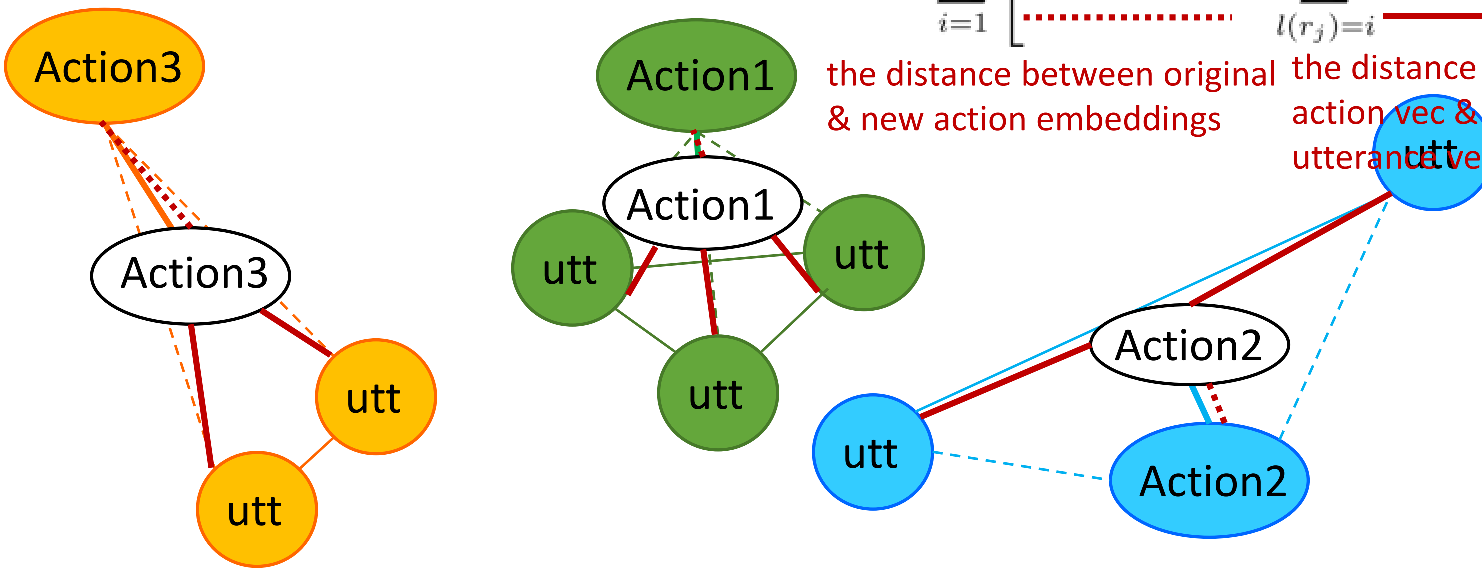
Adapting Action Embeddings

Learning adapted action embeddings by minimizing an objective:

$$\Phi_{\text{act}}(\hat{Q},\hat{R}) = \sum_{i=1}^{n}\left[\alpha_i\|\hat{q}_i - q_i\|^2 + \sum_{l(r_j)=i}\beta_{ij}\|\hat{q}_i - \hat{r}_j\|^2\right]$$

the distance between original & new action embeddings

the distance between new action vec & corresponding utterance vecs



The actionable scores can be measured by the similarity between utterance embeddings and adapted action embeddings.

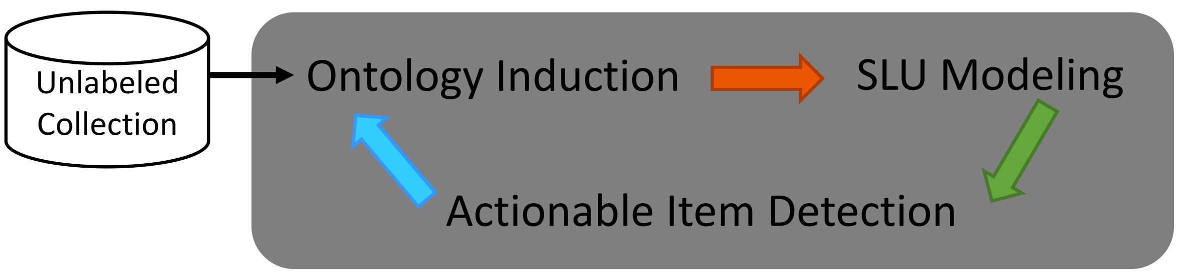
Iterative Ontology Refinement

Actionable information may help refine the induced domain ontology

- Intent: create_single_reminder
  - Higher score → utterance with more core contents
  - Lower score → less important utterance

Utt1	0.7	Receiving Commerce_pay	↑
Utt2	0.2	Capability	↓

$$w'(s) = (1-\alpha)\log \underbrace{f'(s)}_{\text{weighted frequency}} + \alpha \cdot \log \underbrace{h(s)}_{\text{semantic coherence}}$$



The iterative framework can benefit understanding in both human-machine and human-human conversations.

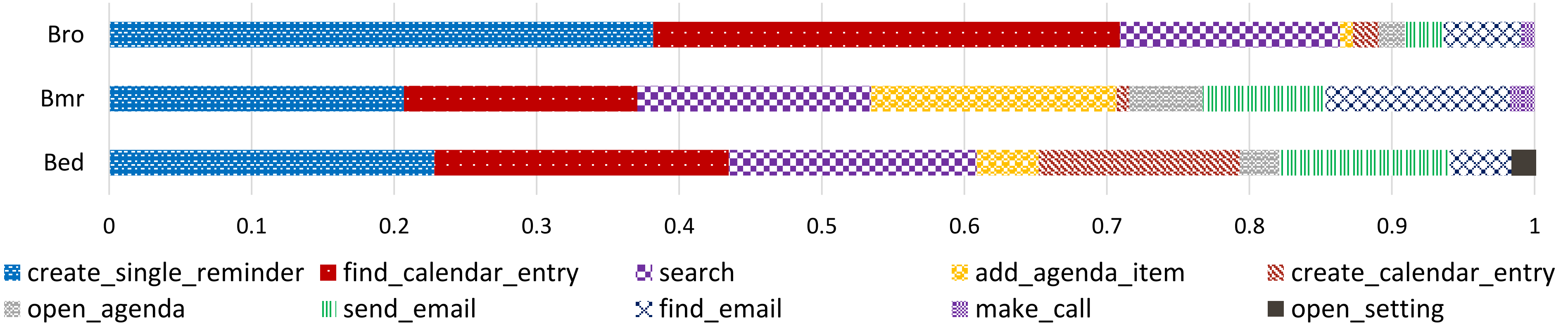
Experiments of H-H SLU

Experiment 1: Actionable Item Detection

Dataset: 22 meetings from the ICSI meeting corpus (3 types of meeting: Bed, Bmr, Bro)

Identified action: *find_calendar_entry, create_calendar_entry, open_agenda, add_agenda_item, create_single_reminder, send_email, find_email, make_call, search, open_setting*

Annotating agreement: Cohen's Kappa = 0.64~0.67



Data & Model Available at http://research.microsoft.com/en-us/projects/meetingunderstanding/

Experiments of H-H SLU

Experiment 1: Actionable Item Detection

Metrics: the average AUC for 10 actions+*others*

trained on Cortana data

then continually trained on meeting data

Approach	Mismatch-CDSSM	Adapt-CDSSM
Original Similarity	49.1	50.4
Similarity based on Adapted Embeddings	55.8 (+13.6%)	**60.1** (+19.2%)

model adaptation

action embedding adaptation

Two adaptation approaches are useful to overcome the genre mismatch.

Experiments of H-H SLU
Experiment 1: Actionable Item Detection

Baselines
- Lexical: ngram
- Semantic: paragraph vector (Le and Mikolov, 2014)

Classifier: SVM with RBF

	Model	AUC (%)
Baseline	N-gram (N=1,2,3)	52.84
	Paragraph Vector (doc2vec)	59.79
Proposed	CDSSM Adapted Vector	**69.27**

The CDSSM semantic features outperform lexical n-grams and paragraph vectors, where about 70% actionable items in meetings can be detected.

Le and Mikolov, "Distributed Representations of Sentences and Documents," in *Proc. of JMLR*, 2014.

Experiments of H-H SLU

Experiment 2: Iterative Ontology Refinement

Dataset: 155 conversations between customers and agents from the MetLife call center; 5,229 automatically segmented utterances (WER = 31.8%)

Annotating agreement of actionable utterances (customer intents or agent actions): Cohen's Kappa = 0.76

Reference Ontology

- Slot: frames selected by annotators
- Structure: slot pairs with dependency relations

FrameNet Coverage: 79.5%

- #additional important concepts: 8
  - cancel, refund, delete, discount, benefit, person, status, care
- #reference slots: 31

Experiments of H-H SLU

Experiment 2: Iterative Ontology Refinement

Metric: AUC for evaluating the ranking lists about slot and slot pairs

Approach	ASR		Transcripts	
	Slot	Structure	Slot	Structure
Baseline: MLE	43.4	11.4	59.5	25.9
Proposed: External Word Vec	**49.6**	**12.8**	**64.7**	**40.2**

The proposed ontology induction significantly improves the baseline in terms of slot and structure performance for multi-domain dialogues.

Experiments of H-H SLU

Experiment 2: Iterative Ontology Refinement

Metric: AUC for evaluating the ranking lists about slot and slot pairs

Approach		ASR		Transcripts	
		Slot	Structure	Slot	Structure
Baseline: MLE		43.4	11.4	59.5	25.9
+ Actionable Score	Proposed	← the estimation of actionable item detection			
	Oracle	← ground truth of actionable utterances (upper bound)			
Proposed: External Word Vec		**49.6**	**12.8**	**64.7**	**40.2**
+ Actionable Score	Proposed				
	Oracle				

The proposed ontology induction significantly improves the baseline in terms of slot and structure performance for <u>multi-domain dialogues</u>.

Experiments of H-H SLU

Experiment 2: Iterative Ontology Refinement

Metric: AUC for evaluating the ranking lists about slot and slot pairs

Approach		ASR		Transcripts	
		Slot	Structure	Slot	Structure
Baseline: MLE		**43.4**	**11.4**	59.5	25.9
+ Actionable Score	Proposed	42.9	11.3		
	Oracle	44.3	12.2		
Proposed: External Word Vec		**49.6**	**12.8**	64.7	40.2
+ Actionable Score	Proposed	49.2	12.8		
	Oracle	48.4	12.9		

Actionable information does not significantly improve ASR results due to high WER.

Experiments of H-H SLU

Experiment 2: Iterative Ontology Refinement

Metric: AUC for evaluating the ranking lists about slot and slot pairs

Approach		ASR		Transcripts	
		Slot	Structure	Slot	Structure
Baseline: MLE		43.4	11.4	59.5	25.9
+ Actionable Score	Proposed	42.9	11.3	**59.8**	**26.6**
	Oracle	44.3	12.2	66.7	37.8
Proposed: External Word Vec		49.6	12.8	64.7	40.2
+ Actionable Score	Proposed	49.2	12.8	**65.0**	**40.5**
	Oracle	48.4	12.9	82.4	56.9

Actionable information significantly improves the performance for transcripts.

The iterative ontology refinement is feasible, and it shows the potential room for improvement.

Outline

- Introduction
- Ontology Induction [ASRU'13, SLT'14a]
- Structure Learning [NAACL-HLT'15]
- Surface Form Derivation [SLT'14b]
- Semantic Decoding [ACL-IJCNLP'15]
- Intent Prediction [SLT'14c, ICMI'15]
- SLU in Human-Human Conversations [ASRU'15]
- **Conclusions & Future Work**

Summary of Contributions

Knowledge Acquisition

- ✓ Ontology Induction → Semantic relations are useful.
- ✓ Structure Learning → Dependency relations are useful.
- ✓ Surface Form Derivation → Web-derived surface forms benefit SLU.

SLU Modeling

- ✓ Semantic Decoding → The MF-SLU decodes semantics.
- ✓ Intent Prediction → The feature-enriched MF-SLU predicts intents.

SLU in Human-Human Conversations

- ✓ CDSSM learns intent embeddings to detect actionable utterances, which may help ontology refinement as an iterative framework.

Conclusions

The dissertation shows the feasibility and the potential for improving *generalization*, *maintenance*, *efficiency*, and *scalability* of SDSs, where the proposed techniques work for both human-machine and human-human conversations.

The proposed **knowledge acquisition** procedure enables systems to automatically produce domain-specific ontologies.

The proposed **MF-SLU** unifies the automatically acquired knowledge, and then allows systems to consider implicit semantics for better understanding.

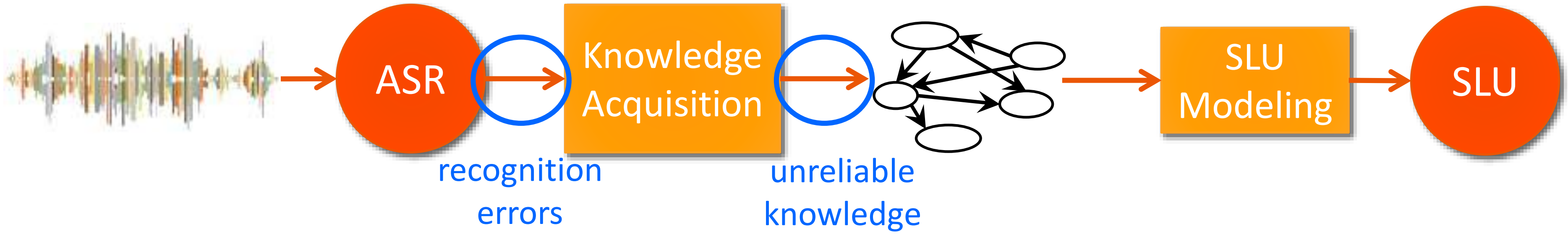
- Better semantic representations for individual utterances
- Better high-level intent prediction about follow-up behaviors

Future Work

Apply the proposed technology to domain discovery

- not covered by the current systems but users are interested in
- guide the next developed domains

Improve the proposed approach by handling the uncertainty



Topic Prediction for ASR Improvement

- Lexicon expansion with potential OOVs
- LM adaptation
- Lattice rescoring

Active Learning for SLU

- w/o labels: data selection, filter uncertain instance
- w/ explicit labels: crowd-sourcing
- w/ implicit labels: successful interactions implies the pseudo labels

Take Home Message

Big Data without annotations is available

Main challenge: how to acquire and organize important knowledge, and further utilize it for applications

Unsupervised or weakly-supervised methods will be the future trend!




THANKS FOR YOUR ATTENTIONS!!

Q & A

THANKS TO MY COMMITTEE MEMBERS FOR THEIR HELPFUL FEEDBACK.