

Accurate Probability Estimation of Hypothesised User Acts for POMDP Approaches to Dialogue Management

Paul A. Crook Oliver Lemon

HCRC/ICCS, School of Informatics
University of Edinburgh, UK



12th annual research colloquium of the special-interest group for computational linguistics in the UK and Ireland
CLUKI 2009, Dublin

Outline

- 1 Background
 - End-to-End Statistical Spoken Dialogue Systems
 - Dialogue Management
 - Learning Frameworks for Statistical Spoken Dialogue Systems
- 2 Experiment
 - Methodology
 - Results

Outline

- 1 Background
 - End-to-End Statistical Spoken Dialogue Systems
 - Dialogue Management
 - Learning Frameworks for Statistical Spoken Dialogue Systems
- 2 Experiment
 - Methodology
 - Results

Context of this Work

- This work should be seen as a contribution to a much large project.
- Aim is to build end-to-end statistical spoken dialogue systems.
- This on going work at the University of Edinburgh.
- Originally funded by the EPSRC (project EP/E019501/1).
- Now with various partners in EU project **CLASSiC**.



UNIVERSITY OF
CAMBRIDGE



UNIVERSITÉ
DE GENÈVE



Supélec

Dialogue Management for Statistical Spoken Dialogue Systems

- My work is specifically focused on Dialogue Management for Statistical Spoken Dialogue Systems.
- So what is a Spoken Dialogue System?
- What is Dialogue Management?
- What are the issues in Dialogue Management?

Dialogue Management for Statistical Spoken Dialogue Systems

- My work is specifically focused on Dialogue Management for Statistical Spoken Dialogue Systems.
- So what is a Spoken Dialogue System?
 - What is Dialogue Management?
 - What are the issues in Dialogue Management?

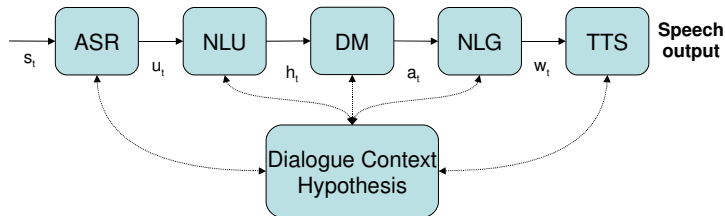
Dialogue Management for Statistical Spoken Dialogue Systems

- My work is specifically focused on Dialogue Management for Statistical Spoken Dialogue Systems.
- So what is a Spoken Dialogue System?
- What is Dialogue Management?
- What are the issues in Dialogue Management?

Dialogue Management for Statistical Spoken Dialogue Systems

- My work is specifically focused on Dialogue Management for Statistical Spoken Dialogue Systems.
- So what is a Spoken Dialogue System?
- What is Dialogue Management?
- What are the issues in Dialogue Management?

What is a Spoken Dialogue System?



Legend:

ASR: Automatic Speech recognition

NLU: Natural Language Understanding

DM: Dialogue Management

NLG: Natural Language Generation

TTS: Text To Speech

s_t : Speech Signal from user

u_t : Utterance Hypotheses

h_t : Conceptual Interpretation Hypotheses

a_t : Action Hypotheses

w_t : Output Hypotheses

Typical Dialogue

System: Hello how may I help?

User: I'm after an expensive French restaurant.

System: You're looking for French food.
In what area of the city do you want to eat?

User: Expensive French in the centre please.

...

Outline

- 1 **Background**
 - End-to-End Statistical Spoken Dialogue Systems
 - **Dialogue Management**
 - Learning Frameworks for Statistical Spoken Dialogue Systems
- 2 **Experiment**
 - Methodology
 - Results

What is Dialogue Management?

Dialogue Management is deciding what the Spoken Dialogue System should do next, *e.g.*

- greet the user,
- request information,
- seek clarification,
- confirm an item,
- search the database,
- present items,
- give up and pass the call to a human,
- close the dialogue,
- etc.

Issues in Dialogue Management?

- Issues with current Dialogue Management:
 - The problem space is very large.
 - Hand-coded solutions are difficult to design and are not guaranteed to be “good”.
 - Systems are fragile, and Users are frustrated!
- Research opportunity:
 - Can we automatically *learn* good solutions in this space?
 - Optimising our dialogue management?

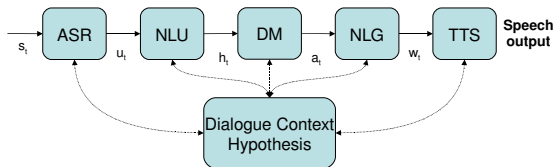
Issues in Dialogue Management?

- Issues with current Dialogue Management:
 - The problem space is very large.
 - Hand-coded solutions are difficult to design and are not guaranteed to be “good”.
 - Systems are fragile, and Users are frustrated!
- Research opportunity:
 - Can we automatically *learn* good solutions in this space?
 - Optimising our dialogue management?

Outline

- 1 Background
 - End-to-End Statistical Spoken Dialogue Systems
 - Dialogue Management
 - Learning Frameworks for Statistical Spoken Dialogue Systems
- 2 Experiment
 - Methodology
 - Results

Single View of a Dialogue



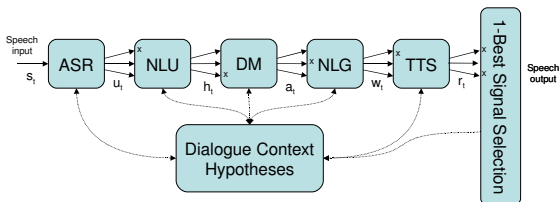
Legend:

ASR: Automatic Speech recognition
NLU: Natural Language Understanding
DM: Dialogue Management
NLG: Natural Language Generation
TTS: Text To Speech

s_t : Speech Signal from user
 u_t : Utterance Hypotheses
 h_t : Conceptual Interpretation Hypotheses
 a_t : Action Hypotheses
 w_t : Output Hypotheses

This view lends itself to learning Dialogue Management as a Markov Decision Process (MDP).

Maintaining Multiple Views of a Dialogue



- Maintaining Multiple Views of the Current Dialogue's State

- s_t : Speech Signal from user
- u_t : Utterance Hypotheses
- h_t : Conceptual Interpretation Hypotheses
- a_t : Action Hypotheses
- w_t : Text output Hypotheses
- r_t : Speech Synthesis Hypotheses
- x : elimination of hypotheses

the Partially Observable Markov Decision Process (POMDP) framework is a natural extension of MDPs that can be applied in this situation

POMDP belief updates

- A POMDP maintain a probability distribution over all possible states of the system, known as its *belief*.
- In this case a distribution is maintained over all possible states of the conversation
 - (a factorised or tree representation is typically used to provide a compact representation... but still not compact enough for learning).
- For each *observation* the POMDP receives it updates its belief space.
 - In the case of Spoken Dialogue System the observation is typically the parsed utterance.
- The standard POMDP belief update equation is of the form: $b'(s') = \sum_{s \in S} P(o|s')P(s'|s, a)b(s)$

POMDP belief updates

- A POMDP maintain a probability distribution over all possible states of the system, known as its *belief*.
- In this case a distribution is maintained over all possible states of the conversation
 - (a factorised or tree representation is typically used to provide a compact representation... but still not compact enough for learning).
- For each *observation* the POMDP receives it updates its belief space.
 - In the case of Spoken Dialogue System the observation is typically the parsed utterance.
- The standard POMDP belief update equation is of the form: $b'(s') = \sum_{s \in S} P(o|s')P(s'|s, a)b(s)$

POMDP belief updates

- A POMDP maintain a probability distribution over all possible states of the system, known as its *belief*.
- In this case a distribution is maintained over all possible states of the conversation
 - (a factorised or tree representation is typically used to provide a compact representation... but still not compact enough for learning).
- For each *observation* the POMDP receives it updates its belief space.
 - In the case of Spoken Dialogue System the observation is typically the parsed utterance.
- The standard POMDP belief update equation is of the form: $b'(s') = \sum_{s \in S} P(o|s')P(s'|s, a)b(s)$

POMDP belief updates

- A POMDP maintain a probability distribution over all possible states of the system, known as its *belief*.
- In this case a distribution is maintained over all possible states of the conversation
 - (a factorised or tree representation is typically used to provide a compact representation... but still not compact enough for learning).
- For each *observation* the POMDP receives it updates its belief space.
 - In the case of Spoken Dialogue System the observation is typically the parsed utterance.
- The standard POMDP belief update equation is of the form: $b'(s') = \sum_{s \in S} P(o|s')P(s'|s, a)b(s)$

POMDP belief updates

- A POMDP maintain a probability distribution over all possible states of the system, known as its *belief*.
- In this case a distribution is maintained over all possible states of the conversation
 - (a factorised or tree representation is typically used to provide a compact representation... but still not compact enough for learning).
- For each *observation* the POMDP receives it updates its belief space.
 - In the case of Spoken Dialogue System the observation is typically the parsed utterance.
- The standard POMDP belief update equation is of the form: $b'(s') = \sum_{s \in S} P(o|s')P(s'|s, a)b(s)$

POMDP belief updates

- A POMDP maintain a probability distribution over all possible states of the system, known as its *belief*.
- In this case a distribution is maintained over all possible states of the conversation
 - (a factorised or tree representation is typically used to provide a compact representation... but still not compact enough for learning).
- For each *observation* the POMDP receives it updates its belief space.
 - In the case of Spoken Dialogue System the observation is typically the parsed utterance.
- The standard POMDP belief update equation is of the form: $b'(s') = \sum_{s \in S} P(o|s')P(s'|s, a)b(s)$

Spoken Dialogue System POMDP belief updates

- For Spoken Dialogue Systems the standard POMDP belief update equation is typically re-factored as:

$$b'(p', a'_u, s'_d) = kP(o'|a'_u)P(a'_u|p', a_m) \sum_{s_d} P(s'_d|p', a'_u, s_d, a_m)P(p'|p)b(p, s_d)$$

- $b'(p, a'_u, s'_d)$ is the updated belief space,
- k is a normalisation constant,
- $P(o'|a'_u)$ is the *observation model*,
- $P(a'_u|p', a_m)$ is the *user action model*,
- $P(s'_d|p', a'_u, s_d, a_m)$ is the dialogue model,
- $P(p'|p)$ and $b(p, s_d)$ together form a belief refinement step.

Spoken Dialogue System POMDP belief updates

- For Spoken Dialogue Systems the standard POMDP belief update equation is typically re-factored as:

$$b'(p', a'_u, s'_d) = kP(o'|a'_u)P(a'_u|p', a_m) \sum_{s_d} P(s'_d|p', a'_u, s_d, a_m)P(p'|p)b(p, s_d)$$

- For this paper the term we are focusing on *observation model*, $P(o'|a'_u)$.
- This term is typically *approximated* by the confidence measures output by the Automatic Speech Recognition (ASR) system.
- It is this approximation that we set out to test.

Spoken Dialogue System POMDP belief updates

- For Spoken Dialogue Systems the standard POMDP belief update equation is typically re-factored as:

$$b'(p', a'_u, s'_d) = kP(o'|a'_u)P(a'_u|p', a_m) \sum_{s_d} P(s'_d|p', a'_u, s_d, a_m)P(p'|p)b(p, s_d)$$

- For this paper the term we are focusing on *observation model*, $P(o'|a'_u)$.
- This term is typically *approximated* by the confidence measures output by the Automatic Speech Recognition (ASR) system.
- It is this approximation that we set out to test.

Spoken Dialogue System POMDP belief updates

- For Spoken Dialogue Systems the standard POMDP belief update equation is typically re-factored as:

$$b'(p', a'_u, s'_d) = kP(o'|a'_u)P(a'_u|p', a_m) \sum_{s_d} P(s'_d|p', a'_u, s_d, a_m)P(p'|p)b(p, s_d)$$

- For this paper the term we are focusing on *observation model*, $P(o'|a'_u)$.
- This term is typically *approximated* by the confidence measures output by the Automatic Speech Recognition (ASR) system.
- It is this approximation that we set out to test.

Outline

- 1 Background
 - End-to-End Statistical Spoken Dialogue Systems
 - Dialogue Management
 - Learning Frameworks for Statistical Spoken Dialogue Systems
- 2 Experiment
 - Methodology
 - Results

Methodology

- Used a transcribed corpus from TownInfo project.
- Parsed the human transcription of each utterance using Spoken Language Understanding (SLU) parsers; GF Parser and Beast Parser.
- The resulting user semantic acts are taken as being the correct interpretation of the user's utterance.
- Against these we compare the outputs of the automatic speech recogniser and parser (ASL-SLU).

Methodology

- Used a transcribed corpus from TownInfo project.
- Parsed the human transcription of each utterance using Spoken Language Understanding (SLU) parsers; GF Parser and Beast Parser.
- The resulting user semantic acts are taken as being the correct interpretation of the user's utterance.
- Against these we compare the outputs of the automatic speech recogniser and parser (ASL-SLU).

Methodology

- Used a transcribed corpus from TownInfo project.
- Parsed the human transcription of each utterance using Spoken Language Understanding (SLU) parsers; GF Parser and Beast Parser.
- The resulting user semantic acts are taken as being the correct interpretation of the user's utterance.
- Against these we compare the outputs of the automatic speech recogniser and parser (ASL-SLU).

Methodology

- Use the Automatic Speech Recogniser HTK/ATK.
- Fed the audio files of each utterance to ATK.
- for each utterance ATK produces an ordered N-best list (where $N=3$) of strings with an associated confidence score.

1st "want an expensive french" 0.92

2nd "want want a french" 0.67

3rd "want an expensive restaurant" 0.32

Methodology

- Use the Automatic Speech Recogniser HTK/ATK.
- Fed the audio files of each utterance to ATK.
- for each utterance ATK produces an ordered N-best list (where $N=3$) of strings with an associated confidence score.

1st "want an expensive french" 0.92

2nd "want want a french" 0.67

3rd "want an expensive restaurant" 0.32

Methodology

- Use the Automatic Speech Recogniser HTK/ATK.
- Fed the audio files of each utterance to ATK.
- for each utterance ATK produces an ordered N-best list (where $N=3$) of strings with an associated confidence score.

1st "want an expensive french" 0.92

2nd "want want a french" 0.67

3rd "want an expensive restaurant" 0.32

Methodology

- Use the Automatic Speech Recogniser HTK/ATK.
- Fed the audio files of each utterance to ATK.
- for each utterance ATK produces an ordered N-best list (where N=3) of strings with an associated confidence score.

1st "want an expensive french" 0.92

2nd "want want a french" 0.67

3rd "want an expensive restaurant" 0.32

Methodology

- The N-best strings are parsed using GF & Beast Parsers to give N-best list of hypothesised user semantic acts.
- The hypothesised user acts binned according to their confidence score.
- And a count made of the number of parsed N-best strings that match with their parsed transcription.
- This gives a rough measure of variation of correctness against ATK confidence score.

Methodology

- The N-best strings are parsed using GF & Beast Parsers to give N-best list of hypothesised user semantic acts.
- The hypothesised user acts binned according to their confidence score.
- And a count made of the number of parsed N-best strings that match with their parsed transcription.
- This gives a rough measure of variation of correctness against ATK confidence score.

Methodology

- The N-best strings are parsed using GF & Beast Parsers to give N-best list of hypothesised user semantic acts.
- The hypothesised user acts binned according to their confidence score.
- And a count made of the number of parsed N-best strings that match with their parsed transcription.
- This gives a rough measure of variation of correctness against ATK confidence score.

Methodology

- The N-best strings are parsed using GF & Beast Parsers to give N-best list of hypothesised user semantic acts.
- The hypothesised user acts binned according to their confidence score.
- And a count made of the number of parsed N-best strings that match with their parsed transcription.
- This gives a rough measure of variation of correctness against ATK confidence score.

Methodology

- ATK can use output two different confidence score measures for each word string;
 - a tradition average word confidence (of all words in the string), and
 - a confusion network based inference evidence score for the complete string.
- We examined both these metrics to see if either provides a good approximation for the observation model.

Methodology

- ATK can use output two different confidence score measures for each word string;
 - a tradition average word confidence (of all words in the string), and
 - a confusion network based inference evidence score for the complete string.
- We examined both these metrics to see if either provides a good approximation for the observation model.

Methodology

- ATK can use output two different confidence score measures for each word string;
 - a tradition average word confidence (of all words in the string), and
 - a confusion network based inference evidence score for the complete string.
- We examined both these metrics to see if either provides a good approximation for the observation model.

Methodology

- ATK can use output two different confidence score measures for each word string;
 - a tradition average word confidence (of all words in the string), and
 - a confusion network based inference evidence score for the complete string.
- We examined both these metrics to see if either provides a good approximation for the observation model.

Outline

- 1 Background
 - End-to-End Statistical Spoken Dialogue Systems
 - Dialogue Management
 - Learning Frameworks for Statistical Spoken Dialogue Systems
- 2 Experiment
 - Methodology
 - Results

Results Breakdown

Results are broken down by:

- Metric: average word confidence, inference evidence score
- The hypothesis position in the N-best list.

As a reminder, the hypothesis we are testing is whether the approximation that: $P(o|a_u) \approx ASR \text{ confidence score}$.

Results Breakdown

Results are broken down by:

- Metric: average word confidence, inference evidence score
- The hypothesis position in the N-best list.

As a reminder, the hypothesis we are testing is whether the approximation that: $P(o|a_v) \approx \text{ASR confidence score}$.

Results Breakdown

Results are broken down by:

- Metric: average word confidence, inference evidence score
- The hypothesis position in the N-best list.

As a reminder, the hypothesis we are testing is whether the approximation that: $P(o|a_v) \approx ASR \text{ confidence score}$.

Results Breakdown

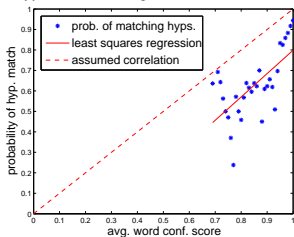
Results are broken down by:

- Metric: average word confidence, inference evidence score
- The hypothesis position in the N-best list.

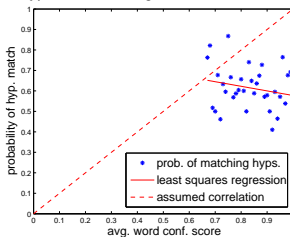
As a reminder, the hypothesis we are testing is whether the approximation that: $P(o|a_U) \approx ASR \text{ confidence score}$.

Average Word Confidence

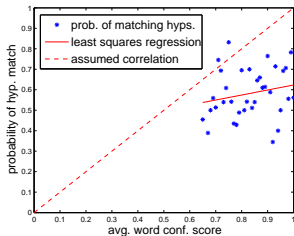
1st hypothesis, average word confidence



2nd hypothesis, average word confidence

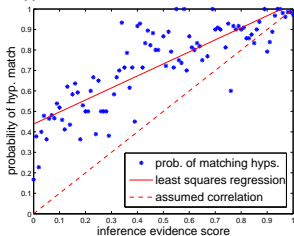


3rd hypothesis, average word confidence

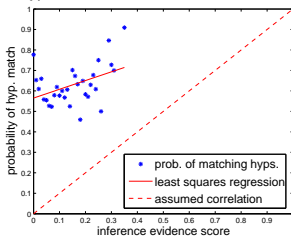


Inference Evidence Score

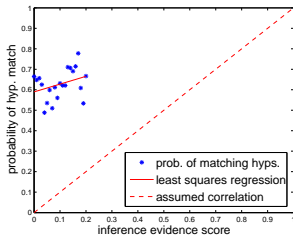
1st hypothesis, inference evidence score



2nd hypothesis, inference evidence score



3rd hypothesis, inference evidence score



Discussion

- Clearly the approximation $P(o|a_u) \approx ASR \text{ confidence}$ score doesn't in general hold.
- However the graphs do suggest a simple linear regression could be used to improve the approximation for the application covered by this corpus.
- With Pearson's correlation suggesting that inference evidence scores should be preferred.

hypothesis number	inference evidence	av. word conf.
1st	0.8439	0.6550
2nd	0.4185	-0.2114
3rd	0.3202	0.2090

Discussion

- Clearly the approximation $P(o|a_u) \approx ASR \text{ confidence}$ score doesn't in general hold.
- However the graphs do suggest a simple linear regression could be used to improve the approximation for the application covered by this corpus.
- With Pearson's correlation suggesting that inference evidence scores should be preferred.

hypothesis number	inference evidence	av. word conf.
1st	0.8439	0.6550
2nd	0.4185	-0.2114
3rd	0.3202	0.2090

Discussion

- Clearly the approximation $P(o|a_u) \approx ASR \text{ confidence}$ score doesn't in general hold.
- However the graphs do suggest a simple linear regression could be used to improve the approximation for the application covered by this corpus.
- With Pearson's correlation suggesting that inference evidence scores should be preferred.

hypothesis number	inference evidence	av. word conf.
1st	0.8439	0.6550
2nd	0.4185	-0.2114
3rd	0.3202	0.2090

Conclusions

- Clearly the approximation $P(o|a_U) \approx ASR \text{ confidence}$ score doesn't in general hold.
- However the approximation can possibly be improved by linear regression.
- The results also suggest inference evidence scores should be preferred for POMDP Spoken Dialogue Systems.
- Future Work
 - Test the effectiveness of the linear remapping with real users.

Conclusions

- Clearly the approximation $P(o|a_U) \approx ASR \text{ confidence}$ score doesn't in general hold.
- However the approximation can possibly be improved by linear regression.
- The results also suggest inference evidence scores should be preferred for POMDP Spoken Dialogue Systems.
- Future Work
 - Test the effectiveness of the linear remapping with real users.

The End

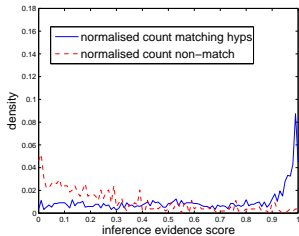
Thank you.

<http://homepages.inf.ed.ac.uk/pacrook>

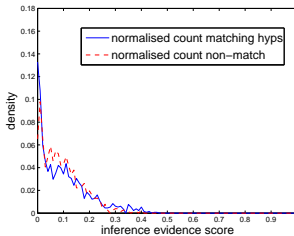
Summed counts of matching and non-matching parsed hypotheses

hypothesis	% matching hyps.	% non-matching hyps.
1st	74	26
2nd	63	37
3rd	62	38

1st hypothesis, inference evidence score



2nd hypothesis, inference evidence score



3rd hypothesis, inference evidence score

