SLT 2014 UNCONFERENCE STYLE SIG MEETINGS PROCEEDINGS

December 7-8, 2014

South Lake Tahoe, NV

Unconference style

Special Interest Group (SIG) meetings

@ SLT 2014

SIG sessions: 9:30-11:00 pm on Sunday, 8:00-9:30 pm and 9:45:11:00 pm on Monday

The concept is fairly simple. An **unconference** is a participant-driven meeting. Typically at an unconference, the agenda is created by the attendees at the beginning of the meeting. Anyone who wants to initiate a discussion on a topic can claim a time and a space once other people sign up for that proposed topic. Unconferences typically feature open discussions rather than having a single speaker at the front of the room giving a talk, although any format is permitted. Group size does not matter, what matters is the discussion content and energy of the group, so even 2 people discussion can create interesting outcomes which can be shared with all the participants. At an unconference, the event lives and dies by the participation of its attendees. They decide what topics will be discussed and they run the meeting. In other words, an unconference has no agenda until the participants create it.

Session proposer sets the floor with a brief introduction and coordinates the discussion in the room which has circular seating style so participants can see each other and a discussion board to outline the session. Session proposer finds someone to take notes for the session and have everyone who comes sign in on the participant form After all notes are submitted by the proposer/s to SIG committee they are compiled into a PDF booklet that will be emailed to participants after the conference. Note taker can use laptop or can take notes using the pen & paper method.

6 easy steps to have unconference style SIG meeting:

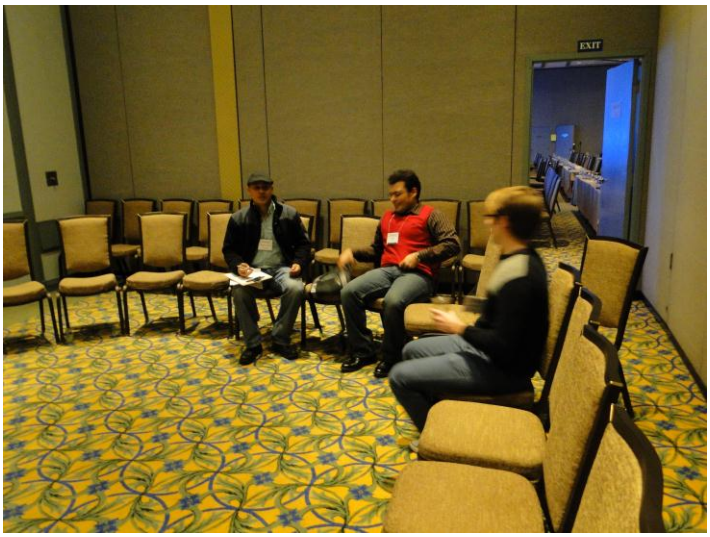
1. Propose a topic and write it down on empty sheet, and post it on the board
2. Waiting period for participants to sign-up (~10-15mins)
3. Ask/find a room or space to host the group discussion
4. Run the meeting and also find a note taker from the participants
5. Compile notes into a summary (info below) and submit to sigs@slt2014.org.
6. Return the filled out SIG form to front-desk

SIG summaries: Submit 1 page slide OR 1 page text with a title and notes & highlights & action items (if there are any) to sigs@slt2014.org. Please make sure to submit your notes **no later than 2pm Tuesday**. During closing ceremony, we will have summary of SIG meetings. Summaries will be shared with all participants after the workshop.

First 15-20 minutes in each SIG session will be used to propose topics and sign-up.

Each night 6 topics were proposed by more than 90 attendees and they were posted on the screen.

Open Space Round Table Style Discussions



Dec 7th Sunday Evening SIG Meetings

[Room names and the topics discussed]

Emerald 1: Dialogue State Tracking Challenge Planning Series #4

Emerald 2: Deeper Understanding

Emerald 3: Trajectory modeling of speech vs DNN.

Emerald 4: Use of SLT in health and education

Emerald 5: Far-field ASR

Emerald 6: If you have 1 year of continuous speech, what would you with it?

Dec 8thMonday Evening SIG Meetings

Emerald 1: Human-in-the-loop approaches to Spoken Language Understanding

Emerald 2: Dialogue State Tracking Challenge (Conference Style)

Emerald 3: Deeper Understanding

Emerald 4: Adaptation of NN? Will it work? What are the alternatives?

Emerald 5: Use of SLT in Health and Education

Emerald 6: How to balance your expectations and the realities of doing a PhD

In total there were 12 SIG meetings for which rooms were allotted both days (6 on each day). However, following meetings were common in both days:

1) Dialogue State Tracking Challenge Planning Series #4
2) Deeper Understanding
3) Use of SLT in health and education

SIG Meetings Summary

S.No.	Topic	Proposer	Note Taker	# of people
1.	Modeling trajectory of speech to try to understand why DNNs work.	Najim Dehak najim@mit.edu	Navid Shokouhi navid.shokouhi@utdallas.edu	8
2.	Discuss plans for dialog state tracking challenge #4	Jason Williams Jason.williams@microsoft.com	Jason Williams Jason.williams@microsoft.com	15
3.	Far field ASR in reverberant and multi speaker conditions.	Sree Hari Krishnan Parthasarthy Sparta@amazon.com	Mahdad Mirsamadi mirsamadi@utdallas.edu	4
4.	Adaptation for neural networks, will it adapt? What is the alternative?	Ozlem Kalinli ozlem.kalinli@ieee.org	Abhinav Misra abhinav.misra@utdallas.edu	13
5.	Deeper models for language understanding and use of A.I. knowledge sources in dialog systems	Gokhan Tur gokhan.tur@ieee.org Tom Kollar, tkollar@apple.com Ron Kaplan RON.KAPLAN@nuance.com	Gokhan Tur gokhan.tur@ieee.org	23
6.	How can SLT technology help in STEM (Math/Science) education and health related issues.	John Hansen john.hansen@utdallas.edu	Masoud Rouhizadeh mrouhizadeh@gmail.com	11
7.	If you had one year of continuously collected speech data, what you would do with it?	Ali Ziaei ali.ziaei@utdallas.edu	Abhijeet Sangwan abhijeet.sangwan@utdallas.edu	6
8.	How to balance your expectations and realities of doing a PhD?	Navid Shokouhi navid.shokouhi@utdallas.edu	Finnian Kelly fpk150030@utdallas.edu	6
Total				**86**

I- Modeling Trajectory of Speech to try to Understand Why DNNs work

Proposer Najim Dehak najim@mit.edu

Note Taker: Navid Shokouhi navid.shokouhi@utdallas.edu

Other Participants: Abhinav Misra, Shabnam Ghaffarzadegan, Ozlem Kalinli, Taufig Hasan, Ananth Shankar, Seyed Hamidreza Mohammadi, Navid Shokouki, Alex Dubindy

Discussion notes:

A number of questions were raised by the presenter (Najeem Dehak):
Is there a way to model the speech trajectory in GMMs?
Would stacking the features be a good idea?
Can we Learn from DNN and map to "classical models" (meaning HMMs and GMMs)?
The use of bottleneck features doesn't answer the question "why is the system working?"
What causes the GMM-HMM that uses bottleneck features give good results?
Can we look at longer segments now that we're not using probabilistic models? [referring the DNNs]
Can DNNs provide a solution to the problems GMMs (HMMs) have in modeling long-term speech trajectories.
DNNs are looking at N frames before and after when they predict a frames' state. This is something that GMMs don't use.

The answer to all these questions could be that we should look into more long-term characteristics, in other words the trajectory of speech.

II- Discuss plans for Dialog State Tracking Challenge #4

Proposer: Jason Williams **Jason.williams@microsoft.com**

Note Taker: Jason Williams **Jason.williams@microsoft.com**

Other Participants: Kai Yu, Kai Sun, Lu Chen, Hang Ren, Weiqin Xu, Yi Ma, Rudolf Kadlec, Ondrej Klejch, Gary Lee, Seokhwan Kim, Lukas Zilka, Dongmo Kim, Milica Gasil, Mathews Hendersan

Discussion Notes:

Seokhwan Kim (A*STAR, Singapore) presented dialog data that his research group has collected, and is interested in making available for a next instance of the Dialog State Tracking Challenge (DSTC). After reviewing the data, the group made a prioritized list of tasks based on this data.

The group consisted of 15-20 participants. Most, but not all, had participated in a previous instance of the dialog state tracking challenge

Seokhwan Kim (A*STAR, Singapore) presented dialog data his research group has collected. The data consists of 35 human-human dialogs, where each is about 30 minutes long. The two participants interacted via Skype and could screen-share. The domain was tourist information for Singapore. 3 people played the part of tour guides, and 35 played tourists. Tourists were not given a specific task to accomplish.

Transcription and speech act labeling has been done already. Seokhwan is open to doing additional labeling, subject to cost. One resource the group felt essential is an ontology of concepts related to the domains studied.

The group made a long list of about 10 tasks, then voted on the most interesting. These were the most popular:

- **State tracking within sub-dialogs**: Filling out a frame of slot-value pairs for a specified region of a dialog, given all dialog history prior to the turn.
- **Speech act prediction**: Predict the speech act of the next turn (either tourist or guide). This can be viewed as policy imitation, where the goal is to imitate the policy of one of the participants.
- **SLU/dialog act tagging**: Given the words, tag turns with speech acts and slots, where slots are text drawn from the utterance.
- **Dialog segmentation**: Segment the dialog into regions, such as "accommodation", "transport information", "attractions", "opening", "closing", etc.
- **End-to-end system**: Using the data and ontology, produce an end-to-end system (playing the part of the guide, or a tourist) capable of conducting new text-based interactions.

All of the tasks seemed feasible to evaluate; all could be evaluated in batch (as has been done in past instances of DSTC), except for the "end-to-end system" task which would require new interactions to evaluate system quality.

It would be desirable if entries were released as services, so that other groups could build on them. For example, in the future, a group attempting to build an end-to-end system could use a state-tracking service, speech act prediction service, etc.

Other tasks which were discussed but which received less interest included: generating output text, summarizing dialogs, selecting the next turn given a set of alternates, co-reference/anaphora resolution, and inferring an onotlogy.

The main next step is that Seokhwan will take this input and refine his proposal for DSTC4, in particular considering what labeling overhead each task would require, and share that updated proposal with the DSTC community.

To receive updates about DSTC4 and participate in its design, join the DSTC4 mailing list by sending an email to: listserv@lists.research.microsoft.com

and put “subscribe DSTC” in the body of the message (without the quotes).

Thanks to Seokhwan for preparing a talk about the A*STAR data and making an initial proposal for a task. Thanks also to the SLT organizers.

III- Far field ASR in Reverberant and Multi Speaker Conditions.

Proposer: Sree Hari Krishnan Parthasarthy Sparta@amazon.com

Note Taker: Mahdad Mirsamadi mirsamadi@utdallas.edu

Other Participants: Shk Parthasarathi, Matt Mirsamadi, Pancha Sankaran, Ananth Shankar

Discussion Notes:

Discussed following items.

- Closely-spaced arrays vs distributed arrays.
- Does feature cleaning help with DNNs?
- Why do we think we understand GMMs better than DNNs?
- Do we need pre-training?

IV- Adaptation for Neural Networks: will it adapt? What is the alternative?

Proposer: Ozlem Kalinli ozlem.kalinli@ieee.org

Note Taker: Abhinav Misra abhinav.misra@utdallas.edu

Other Participants: Fawel Swietojanski, Mohammed Abdelwahab, Najim Dehak, Yajie Miao, Abhinav Miasra, Matt Mirsamadi, Seyed Hamidreza Mohammadi, Weiran Wang, Sachin Kajarekav, Emmanuel Dveous, Panchi Sankaran, Xavier Menendaz Pital, Thomas Schaaf

Discussion Notes:

The group met to discuss adaptation of deep neural networks. There were experts from different backgrounds including people working on noise robustness, speaker adaptation, speaker ID/verification, deep neural networks. Initially, the obvious solutions like feature and speech enhancement for noisy speech and multi-condition training are mentioned. The work on adaptation of neural networks to noise is very limited; however, there is more initiative on the speaker adaptation of neural networks. Fawel described some of the work on adapting DNNs to speakers. A discussion took place around whether similar methods could be applied for noise adaptation. Also, which layers of DNNs can be more appropriate for noise and speaker adaptation is speculated. Najem mentioned the SIG meeting took place last night which was on trajectory modeling of speech for better understand why DNN's work, which can enlighten us on adaptation as well.

Some other highlights of the meeting are listed below:

- Training a neural network on a noisy speech and a neural network on enhanced speech and then decoding it dramatically enhances the performance.
- But de-noising speech doesn't help speaker recognition, as it removes parts of the spectrum that might contain useful speaker information.
- GMMs are generative but neural nets learn boundaries. So, the question is how to shift boundaries to accommodate noise. This question might solve the problem of adapting neural nets to noisy data.
- In adapting neural nets, there is also one problem of speed in online decoding.
- DNNs in Speaker ID are restricted to the extraction of sufficient statistics. It has replaced UBM to give a more structured UBM as it improves performance.
- Inspired from some of speaker adaptation work, instead of adapting the whole DNN for noise may be just adapt the bottom layer.

In summary, it was agreed that the work on adaptation of DNN is limited and deeper understanding of why DNN's work may be needed to succeed on adaptation.

V- Deeper models for language understanding and use of A.I. knowledge sources in dialog systems

Proposers: Gokhan Tur gokhan.tur@ieee.org and Tom Kollar, tkollar@apple.com and Ron Kaplan RON.KAPLAN@nuance.com

Note Taker: Gokhan Tur gokhan.tur@ieee.org

Other Participants: Scott Cyphers, Mandy Korpusik, Nabal Naraula, Kadri Hacioglu, Trung Bui, Ali Orkan Bayer, Chuck Wooters, Pascale Fung, Ron Kaplan, Alex Dubinsky, Yang Liu, Asli Celikyilmaz, Yun-Nung (Vivian) Chen, Murat Akbacak, Thomas Kollar, Dilek Hakkani-Tur, Frederic Bechet, Nicolas Scheffer, Stanley Peters, Ryn , Jung Yun Seo

Discussion Notes:

This SIG is organized as a roundtable discussion bringing together experts with different backgrounds working on language understanding. Robust language understanding has the potential to revolutionize our interactions with computers. Apple's Siri, Microsoft's Cortana and Google Now have set a precedent for personalized, robust speech recognition and language understanding. Despite this, there remain challenges to creating intelligent agents that operate in the presence of imperfect information, with speech recognition errors / noise and that can handle the depth and breadth of human language and knowledge. Current research has shown significant progress in targeted semantic template filling, compositional semantic parsing, discourse understanding, grounded language understanding, and interactive learning.

During the warm up session on the first day we have discussed to identify key discussion topics for the main discussion. We have focused on 4 areas:

1. System knowing more information, static or contextual about the user

2. Users teaching the system or giving immediate feedback for better performance

3. System providing immediate feedback explaining what it understood and why

4. Compositional semantic parsing with discourse

The bottom line for all these 4 discussion points was the semantic representation as an extension for well-known targeted understanding. Ron gave this example of planning a "romantic evening". How would one integrate this concept in the semantic representation of a system. One approach is taking the AI stand, and try to learn and encode these. Another approach would be do not define anything, let the system learn using data. Pascale Fung emphasized the role of the end goal, instead of building a generic semantic representation. Ron Kaplan has also emphasized inferences coming with these semantic knowledge graph, and he made a distinction between structured knowledge bases, mostly used for entities. We have discussed a little bit about a new task, the Nuance Winograd Challenge, along with Watson question answering task.

In the second day, the participation was maximum and discussed in a full room about these 4 topics. Giuseppe Riccardi mentioned the explicit benefits of using a standard representation like FrameNet for the whole community. Roberto Pieraccini told the ATIS days, where they struggled to define a semantic representation and favored for a simpler English translation which can be converted into a function or SQL. Dilek Hakkani-Tur also emphasized the role of the backend database or knowledge graph to match instead of deriving some representation. Stanley Peters mentioned that these are important problems for ASR and MT but critical for NLU.

In the last part we discussed users teaching the system. Tom Kollar mentioned that there a non-toy system in the robotics community, where we can adopt some ideas.

For compositional semantics we discussed briefly the trade off between answering long queries which require compositional semantics and short natural human/human-like interactions.

In summary, we believe this is the first attempt towards bringing experts together discussing these key aspects for language understanding and we hope to have a more formal and structured version soon.

VI- How can SLT technology help in STEM (Math/Science) education and health related issues?

Proposers: John Hansen john.hansen@utdallas.edu Masoud Rouhizadeh mrouhizadeh@gmail.com

Note Taker: Masoud Rouhizadeh mrouhizadeh@gmail.com

Other Participants: Mahsa Elyasi, Ruxin Chen, Finnian Kelly, Ali Ziaei, Hussnain Ali, Seyed Hamidreza Mohammadi, Masoud Rouhizadih, Najim Dehak, Diego Giuliahi

Discussion Notes:

Stem Education:

- Peer lab discussion group
- Word count estimation / turn taking to track discussions.

Health:

- Alzheimer, Parkinson, Autism
- Schedule appointment
- Keep tracking / monitor health

Over recent years, education has become a worthwhile application for speech and language technology (SLT). SLT for education can be used in first and second language learning and acquisition, spoken dialogue systems for education, intelligent tutoring systems, tutor assessment, etc.

One application of SLT can be the evaluation and improvement of the student's performance in Peer-Led Team Learning (PLTL) groups. PLTL are a student-centered active-learning pedagogies commonly used in STEM education. They supplement the huge classroom lectures with group work sessions including 4-8 students from heterogeneous grade levels. It has been shown that the student's level of engagement in PLTL has a meaningful correlation to their performance in the course. SLT can provide an objective way to measure each individual's engagement level by measuring his/her turns and utterances, number of words, etc.

Mental health problems are one of the most difficult challenges in every society. As an example 1 in every 68 children is diagnosed with autism spectrum disorder. Speech and language features play a very important role in diagnosis for mental heath professionals, both in the standard neuropsychological assessment instrument, as well as the clinical impressions of the health expert based on their conversations with the patient.

Many individuals (or their families) might not even be aware of their existing mental health problems due to lack of access to qualified clinician for assessment. In addition, those who are aware of their mental problems may have difficulty in accessing professional treatment due to lack of adequate insurance or financial resources, living in rural areas, cultural and language barriers, etc. Even if the patients have an adequate access to professional help, the limited-time clinical visits might not provide the "full picture" of the mental health issues of the individual.

Speech and language technology can have a huge impact in both education and health domain:

1) Developing inexpensive education evaluation and mental health screening measures that can be much more widely used.

2) Providing objective and consistent evaluation measures as apposed to subjective human impressions.

3) Building conversation agents/assistant to help in the learning/treatment process.

4-a) Online and real-time monitoring of the students' interactions with other students (e.g. in the PLTL groups) or with the instructor (e.g. in the regular classroom environment) in order to help 1) evaluating the effectiveness of the education, 2) discovering new speech and language related features which do not exist in the current assessment instrument.

4-b) Online and real-time monitoring of the patient's behaviors (e.g. at home) to help 1) evaluating the effectiveness of the treatment process, and 2) discovering new speech and language related features which do not exist in the current (limited-time) assessment instrument.

VII- If you had one year of continuously collected speech data, what you would do with it?

Proposer: Ali Ziaei ali.ziaei@utdallas.edu

Note Taker: Abhijeet Sangwan abhijeet.sangwan@utdallas.edu

Other Participants: Lukas Zilka, Ondrej Klejch, Abhijeet Sangwan, Ali Ziaei, Finnan Kelly, Meysam Asgari

Discussion Notes:

Applications:

- 100 students wear one recorder.
- 30 sec. of data every 12 min. every day 8 a.m. to 10 p.m.
- Depression.
- Annotation of location for each sample.
- Activity annotation.
- What are people talking about?
- Measure depression. 85% accuracy.
- How to remove dependency on continuous annotation.
- Context.
- Quality of life, Behavior, performance.
- Diagnosis-Alzheimer.

Privacy:

- Similar problem in biometric world.
- voice conversion system.

The SIG discussed applications and technical challenges faced when working with long duration audio. SIG members shared their research experience which ranged from applications in health (detecting depression in students), education (peer led team learning programs) and life-logging (Prof-Life-Log). The SIG members identified annotations/transcription as a major challenge, and discussed new ideas that could help solve this problem. Members also shared recent advancements made in speech processing technology that addresses some unique challenges faced in long duration audio (for example, long periods of sparse or non-speech,

widely ranging environmental noise conditions, etc.) Furthermore, the members also agreed that continued research on long duration audio can yield new applications in quality of life, human behavior and performance, health and education domains. The issues of privacy was also discussed and the members discussed the scope of the problem and possible solutions.

Overall, the members were excited to be a part of the 'unconference style' SIG meetings at SLT 2014. They thought it was a good platform for engaging in productive discussions, network with other conference attendees, share ideas, know-how and knowledge in general. They are eager to see 'unconference style' become standard at future events.

VIII- How to balance your expectations and realities of doing a PhD?

Proposer: Navid Shokouhi navid.shokouhi@utdallas.edu

Note Taker: Finnian Kelly fpk150030@utdallas.edu

Other Participants: Jeesoo Bang, Navid Shokouhi, Qian Zhang, Murat Akbacak, Stephanie Pancoast

Discussion Notes:

Some challenges/difficulties new PhD students face:

- Integrating with their research group
- 'Impostor syndrome' – i.e. "I am not good enough to be here!"
- Putting small research-related problems in perspective
- Department level assignment of tutors like professional mentors.
- Conference student sessions.
- Importance of supervisor.

A positive working environment is very important. Some small initiatives could have a positive impact on PhD student's outlook

- Designating a senior PhD student as a social coordinator for a research group (in the case where this does not happen organically)
- Assigning a senior PhD student as a mentor to each incoming PhD student (not necessarily in the same field). They would offer informal advice on all personal, social and research related concerns that the new student may have. They would meet regularly, particularly in the first few months. This initiative could be department led, with meetings being mandatory for both parties. This would be particularly beneficial for small/dispersed research groups.
- At conferences, there should be informal sessions just for PhD students to voice any thoughts or concerns they may have about the PhD process.