

Preworkshop Activities v.2007

Sanjeev Khudanpur

Department of Electrical and Computer Engineering

and

Center for Language and Speech Processing

Johns Hopkins University

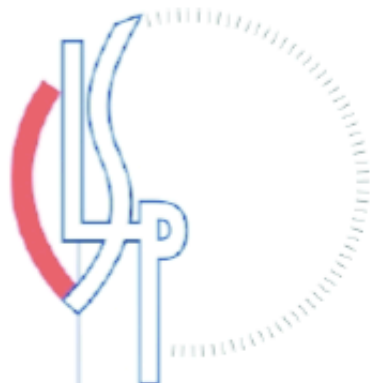
The 2007 Summer Workshop Opening Day Presentations

July 16, 2007.

The Annual CLSP Workshop Cycle

The "summer" workshop is a year-round activity!

September	October	November	December
Solicit proposals Submit final reports	Proposal deadline Preliminary screening Review mtg invitations	Peer-Review Meeting Select Team Leaders Form skeletal teams	\$\$ from NSF and DoD Advertise for UGrads (deadline Feb 2007)
January	February	March	April
Finalize Sr. members Data requirements Computing requirement	Select Grad students Data contracts given Order computers	Select Undergrads 1st team meeting Pilot data arrives	Reconfigure computers Refine data specs Perform pilot expts
May	June	July	August
2nd team meeting More data arrives Start baseline expts	Workshop lab set-up All data arrives Finish baseline expts	Summer School Workshop Begins Weekly team updates	Workshop Ends Write final reports Call for proposals



The Center for Language and Speech Processing
at Johns Hopkins University invites one page research proposals for a
Summer Workshop on Language Engineering,
to be held in Baltimore, MD, USA,
July 9 to August 17, 2007.

CALL FOR PROPOSALS

Deadline: **Wednesday, October 18, 2006.**

You may already know about these summer workshops, which we have hosted since 1995. This year, we have identified specific research topics on which progress is desired. We are therefore soliciting research proposals (suitable for a six week team exploration) in the following research areas:

- **MULTIDOCUMENT, MULTILINGUAL ENTITY DISAMBIGUATION:** *Content extraction is an extremely important area of research. Entity disambiguation – determining whether two entity mentions have the same referent – is a very important sub-problem. Disambiguation is challenging within one document; it is even more challenging and more important across documents, especially in multilingual and/or multi-genre collections. The ultimate goal is to have automatic methods to identify all the unique entities in a collection of documents of varied types in several languages, and to associate each such entity with all its mentions, including nicknames, name variations, misspellings, translations, transliterations, and anaphors. Algorithmic solutions to these or a subset of these problem are of significant interest.*
- **AUTOMATIC ADAPTATION OF SPEECH TECHNOLOGY TO NEW DOMAINS:** *A recurring problem in creating and refining speech technology is the time and effort required for data annotation. Another is the need to update models as channels, speakers, and vocabularies change over time. It would be extremely helpful to have algorithms that can automatically adapt speech processing systems as their input data evolves. It would be very helpful to study techniques for making systems trained on large amounts (e.g., thousands of hours) of speech data from a variety of domains (different subject matters, styles, speakers, and channels) perform nearly as well in new or unfamiliar domains as they do in the seen domains, using only tens of hours of annotated data from the new domains. Techniques that use large amounts of un-annotated data from new domains would also be appropriate.*
- **SOCIAL NETWORKS AND LANGUAGE:** *Identifying groups and social roles of individuals from the frequency and linguistic content of communications poses questions at the intersection of social network theory, graph theory and natural language processing. Previous work on the topology of social networks unveiled surprising characteristics of human networks (e.g., Milgram's experiments of the 60's) and of the connectivity of websites on the Internet. It would be helpful to understand how these theories apply to on-line communities associated with blogs, chat-rooms, instant messaging, etc. Algorithmic solutions that accurately identify groups in these communities are particularly desired.*

New WS'07 Team Selection Process

- Discussed workshop **themes** with sponsors in May 2006
 - Solicited proposals from suggested “thought leaders”
 - Received several thoughts, but few commitments for “team leaders”
- Open call for proposals issued September 2006
 - Electronically advertised to 100’s of researchers
 - Advertised at ACL, HLT-EMNLP, ICASSP, InterSpeech, **GALE**, ...
 - Government program managers recruited to advertise the CFP
- 12 proposals received by October 18, 2006
- 2-day peer review panel met November 3-5, 2006
 - Authors, peers, government PMs and JHU faculty discussed, revised and voted on proposals (STV)

Proposals Considered for WS'06

Presenter (Affiliation)	Proposal Title
Sherri Condon (MITRE)	Transliteration of Names for Entity Disambiguation
David Day (MITRE)	Exploiting Wikipedia for Entity Disambiguation
Louise Guthrie (Sheffield)	Multidocument Multilingual Entity Disambiguation (withdrawn)
Eric Hughes (MITRE)	Representing Uncertainty in Entity Extraction
Massimo Poesio (Essex)	Entity Tracking; Pushing the State of the Art
Richard Sproat (Illinois)	Multilingual Multidocument Entity Disambiguation
Hynek Hermansky (IDIAP)	Identification and Recognition of Unexpected Lexical Terms
Chin-hui Lee (GA Tech)	Acoustic Model Adaptation from Universal Phone Models
Herve Bouillard (IDIAP)	Joint Speaker Identification and Social Network Inference
Owen Rambow (Columbia)	Social Networks and Language
David Jensen (U Mass)	Learned Models of Graph- and Network-Structure Indices
Robert Warren (Waterloo)	Social Networks and Language

Attendees of the Nov 11-13 Meeting

Herve Bourlard (IDIAP)	Aaron harnly (Columbia)	Owen Rambow (Columbia)	Goeffrey Zweig (MSR)
Sherri Condon (MITRE)	Hynek Hermansky (IDIAP)	Astrid S-Neilson (ONR)	
David Day (MITRE)	Frederick Jelinek (JHU)	Patrick Schone (DoD)	
Jason Eisner (JHU)	David Jensen (UMass)	Ted Senator (ex DARPA)	
Jack Godfrey (DoD)	Damianos Karakos (JHU)	Richard Sproat (Illinois)	
Allen Gorin (DoD)	Sanjeev Khudanpur (JHU)	Robert Warren (W'loo)	
Louise Guthrie (Sheffield)	Tatiana Korelsky (NSF)	Charles Wayne (DoD)	
Jan Hajic (Charles Univ)	Chin-hui Lee (GA Tech)	Barb Wheatley (DoD)	
Keith Hall (JHU)	Massimo Poesio (Essex)	David Yarowsky (JHU)	

Panel Composition

25% Government

14% Industry / Labs

39% Academia (non-JHU)

21% JHU Faculty


The WS'07 Teams & Team Leaders

- Recovery from Model Inconsistency in Multilingual Speech Recognition
 - Hynek **Hermansky** (IDIAP), Lukas **Burget** (Brno), Jon **Nedel** (DoD), Chin-Hui **Lee** (GA Tech), Geoffrey **Zweig** (MSR), Haizhou **Li** (IIR), Sanjeev **Khudanpur** (JHU)
 - **Grads**: Petr **Schwarz** (Brno), Pavel **Matejka** (Brno), Chris **White** (JHU), Ariya **Rastrow** (JHU), Rong **Tong** (NUS)
 - **Undergrads**: Sally **Issacoff** (Michigan), Mirko **Hanemann** (Magdeberg), Puneet **Sahani** (NCE, Delhi)
- Exploiting Lexical and Encyclopedic Resources for Entity Disambiguation
 - Massimo **Poesio** (Essex), David **Day** (MITRE), Ron **Artstein** (Essex), Alessandro **Moschitti** (Rome), Jason **Duncan** (DoD), Xiaofeng **Yang** (IIR)
 - **Grads**: Simone **Ponzetto** (EML), Yannick **Verseley** (Tubingen), Michael **Wick** (U Mass), Robert **Hall** (U Mass), Jason **Smith** (JHU)
 - **Undergrads**: Alan **Jern** (UCLA), Brett **Schwom** (NYU), Vladimir **Eidelman** (Columbia)

The JHU-NAACL Summer School on Human Language Technology

- A two week tutorial on all aspects of human language technology by expert academics and practitioners
 - ASR, Dialog Systems, MT, Machine Learning, NLP, IR, ...
- Started in 1998 to bring the undergraduate participants up to speed on human language technology
 - Attended by DoD and other local participants looking for a short comprehensive overview of language technology
 - Additional participants starting 2002 nominated by NAACL
- Attendance routinely exceeds 30 students
 - 25% undergraduates
 - 65% graduate students (more than half are "beginners")
 - 10% researchers (mostly DoD)

2007 JHU/NAACL Summer School

Mon July 2	Tue July 3	Wed July 4	Thu July 5	Fri July 6
Core NLP	Intro to IR	Holiday	Intro to IE	Intro to ML
Jason Eisner	Jim Mayfield		Radu Florian	Laurent Younes
Eugene Charniak	Ellen Voorhees		Regina Barzilay	Ben Tasker
Jason Eisner	David Yarowsky		Radu Florian	Karen Livescu
Mon July 9	Tue July 10	Wed July 11	Thu July 12	Fri July 13
Intro to ASR	Applications	Intro to MT	NE Team	ASR Team
Paul Bamberg	Allison Powell	Chris Callison-Burch	Massimo Poesio	Lukas Burget
			Simone Ponzetto	Wayne Ward
James Glass	Tim Paek	David Yarowsky	Poesio, Ponzetto and Michael Wick	Hermanskyites
	R. Pieraccini			

The 2007 Workshop Calendar

	Monday	Tuesday	Wednesday	Thursday	Friday
July	16	17	18	19	20
	Steering Meeting		UG Lunch and Group Dinner	Tobias Scheffer	Ice-cream social
	23	24	25	26	27
	Steering Meeting		Rene Vidal	UG Lunch and Group Dinner	Ice-cream social
	30	31	1	2	3
	Steering Meeting		Kevin Cohen	UG Lunch and Group Dinner	Ice-cream social
August	6	7	8	9	10
	Steering Meeting		Jont Allen	UG Lunch and Group Dinner	Ice-cream social
	13	14	15	16	17
	Steering Meeting		Corinna Cortes	UG Lunch and Group Dinner	Ice-cream social
	20	21	22	23	24
	Steering Meeting		Closing Presentations		Farewell Lunch

Encourage Attendance @ WS'07

Events: Invite Your Friends!

- All workshop presentations are open to the public
- If possible, tell us in advance if you plan to attend, to help us with the logistics: clsp@clsp.jhu.edu
- Contact CLSP staff for local assistance in Baltimore
- Mark **August 22**, 9:00 AM to 4:00 PM, on your calendars
-- final presentations of WS'07