The 2nd International Symposium on Language Resources and Intelligence

Penn Wharton China Center, Beijing
(in cooperation with Beijing Language and Culture University)

Monday, December 17th (Day Two)

8:30 – 9:00 – Introductions

Mark Liberman, University of Pennsylvania
Yuming Li, Beijing Advanced Innovation Center for Language Resources

9:00 – 9:30 – Christopher Cieri, Linguistic Data Consortium

"NIEUW: Infrastructure for Eliciting Language and Linguistic Judgments in Multiple Languages and Situations"

Abstract: Despite the efforts of numerous data centers, research labs, national corpus initiatives and programs worldwide, there are still far fewer language resources available than are needed to document and support research and technology development for the world's languages, or even the small subset currently receiving intensive attention. The gap between demand and supply results from the sheer number of languages and variety of datasets needed relative to the human and financial capital devoted to plugging the gaps. The NIEUW project complements project-oriented data collection efforts which use monetary compensation and their principal incentive. NIEUW attempts to address the paucity of language resources by offering novel, renewable incentives, such as entertainment and education, to potential contributors and designing workflows for collecting language data and judgments that match the skills of the workforces attracted by such incentives. NIEUW is developing software infrastructure to support the creation of activities that solicit data contributions and judgments from people with wide range of motivations and skills. These activities are gathered into portals to attract clusters of potential contributors with similar motivations. By minimizing dependence on direct project funding, NIEUW reduces constraints on language, volume, and timeline. Relevant examples of NIEUW activities include games that elicit language and dialect identification judgments and citizen science activities that collect lexical items and relations, structured translations, and prompted speech including picture descriptions.

Bio: Christopher Cieri's interests lie at the intersection of language, large data, and computation. He received his PhD in Linguistics from the University of Pennsylvania where he focused on sociolinguistics, language contact, phonetics, phonology and morphology. He has worked since 1983 applying technology to linguistic analysis and language teaching using data sets that are too large to process with purely human effort. Cieri became the Executive Director of the Linguistic Data Consortium in 1998 and has since been responsible for overseeing all aspects of the Consortium's operations including the publication of over 500 data sets and the management of many sponsored programs. His current work focuses on the science of linguistic annotation and the analysis of conversational data, most recently in identifying linguistic features to correlate with clinical diagnostic categories and in the incentives that motivate people to contribute linguistic data and judgments and the tools and workflows required by such providers.

9:30 – 10:00 – Jiahong Yuan, Linguistic Data Consortium / LAIX Inc.

"Classification of Chinese dialect regions from L2 English speech"

Abstract: Generally speaking, foreign accent in a second language (L2) results from the interference of the learner's first language (L1). Compared to classification of native languages from L2 speech, classification of Chinese dialect regions is more challenging because of the impact of Mandarin. In this talk, we present an effort to classify Chinese speakers' L1 dialect regions from their L2 English speech, based on the analysis of 1600 hours of speech data collected from a mobile app.

Bio: Jiahong Yuan is associate director of speech research at the Linguistic Data Consortium, currently on leave and working with Laix Inc. His research interests are speech prosody, corpus phonetics, and the integration of speech technology and phonetics research.

10:00 – 10:30 – Jun Du, University of Science and Technology of China

"Speech Processing in Realistic Environments: From Speaker Diarization to Speech Recognition"

Abstract: Recently, with the rapid development of speech-enabled applications, the new research trend of speech signal processing is to enhance the corrupted speech in multi-talker conversations under realistic adverse environments. This is quite challenging as it requires the joint modeling of multiple irrelevant factors such as the background noises, reverberations, and interferring speakers. In this talk, I will take the newly launched DIHARD-I Challenge for speaker diarization and CHiME-5 Challenge for speech recognition as two examples to introduce our recent progress and summarize new challenges.

Bio: Jun Du received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. From July 2009 to June 2010, he worked with iFlytek Research. From July 2010 to January 2013, he joined Microsoft Research Asia as an associate researcher. Since February 2013, hehas been with USTC as an associate professor. His research interests include speech signal processingand pattern recognition. He is the Associate Editor of IEEE/ACM Transactions on Audio, Speech and Language Processing (2018.07-2021.07).

10:30 – 10:45 – Coffee/Tea Break

10:45 – 11:15 – Biyan Yu & Yue Jiang, Xi'an Jiaotong University

“Probability distribution of syntactic divergences of determiner his-(adjective)-noun structure in English-to-Chinese Translation”

Abstract. Studies of translation divergences reveal that lexical divergences in English-to-Chinese translation some regularities which can be modeled by a probability distribution. Besides lexical divergences, we hypothesize that there might be a great number of syntactic divergences since Chinese and English belong to two typological divergent languages. In this study, we investigated whether the Chinese translation of English determiner *his*-(adjective)-noun structure (DNS (*his*) hereafter) diversifies in syntactic construction and whether the distribution of diversified Chinese translations conforms to the regularity of diversification process introduced by Altmann (2005). It is found that the twenty-two Chinese syntactic constructions corresponding to one single English DNS (*his*) do form a decreasing rank-frequency distribution, which is in comfortable agreement with the usual Zipf-Alekseev function. This diversification process may be ascribed to both linguistic and cultural factors such as contextual factors, translator’s subjectivity, functional equivalents, and the tendency for minimal effort in language coding and decoding during translating process.

Bio: Biyan Yu received the PhD from Xi’an Jiaotong University in 2018, and is now a teaching assistant at Chang An University. Dr. Yu’s research interests include quantitative linguistics and translation studies.

Yue Jiang received the PhD from the PLA Institute of Foreign Languages in 1988, and is now professor of ESL and Linguistics in the School of Foreign Studies at Xi’an Jiaotong University.

11:15 – 11:45 – Hui Feng, Tianjin University

"Perception and production of English lexical stress by Uyhgur and Tibetan speakers"

Abstract. This study focuses on the production of English lexical stress by native speakers of Tibetan and Uyghur, and the factors that may affect stress assignment. Thirty subjects in their twenties participated, with 10 native speakers (gender balanced) for each language, i.e. native speakers of Uyghur (NSUs), native speakers of Tibetan (NSTs) and native speakers of American English (NSAs). A total of 4,000 tokens are collected, judged and analyzed. Results indicate that: (1) Consistent with the prediction of Stress Typology Model, less negative transfer has been observed in NSTs than in NSUs in stress production; (2) Compared with NSAs, NSUs and NSTs employ different acoustic features when assigning stress; (3) Stress positions affect the accuracy of stress production by NSUs, and also the acoustic features of NSTs and NSUs when assigning stress, and a speech-final lengthening effect is observed; (4) Syllable structures have little effect on the accuracy of stress production.

Bio: Hui Feng received her B.A. degree, M.A. degree and Ph. D. degree from Nankai University, China, in 1995, 2002, and 2007 respectively. She joined the Department of English at Tianjin University as a research associate in 1995 and was a lecturer from 1995 to 2007. Since 2007 she has been an associate professor in the School of Foreign Language and Literature, Tianjin University, China. She did her post-doc research in the Department of Linguistics of the University of Arizona, USA, in 2011-2012. She was a visiting scholar in the Department of Linguistics in the University of Pennsylvania in 2017-2018. Her research interests include experimental acoustics, second language acquisition and sociolinguistics.

11:45 – 12:15 – Kenneth Church, Baidu Research

"Chasing the unexplained variance"

Abstract: Deep nets are hot. We will start by giving some demos (shameless plugs) of deep nets at Baidu. Could there be too much excitement? Too much excitement has led to AI Winters in the past. Minsky and Chomsky started out their careers in the 1950s questioning a number of machine learning methods that have since regained popularity. Their arguments are largely rejected these days as too negative (ngrams can't do this and nets can't do that). But their work led to constructive suggestions such as the Chomsky Hierarchy. The deep nets literature could benefit from an organization of the literature like the Chomsky Hierarchy. Chomsky's interest in the hierarchy involved long distanced dependencies. In addition to the kinds of dependencies that he talked about (sentence internal constraints such as agreement and wh-movement), there are also lots of important constraints across sentences (coherence, topic, genre, author). It is said that modern nets (such as LSTMs) improve over previous method (ngrams) because they do a better job of capturing long distance dependencies (including both the syntactic constraints that Chomsky was interested in, as well as document-level constraints such as coherence). It is common in language modeling to condition on what we can afford to condition on (such as the recent past), and assume that everything else doesn't matter. LSTMs (and topic models) can be viewed as variants, where we condition on different things, but at the end of the day, we introduce an assumption that everything else doesn't matter. The unexplained variance is typically much larger than what would be expected under standard independence assumptions (such as Poissons). LSTMs typically scramble sentences because they don't capture long-distance dependences across sentences. Obviously we can't capture everything, but it would be better to estimate the unexplained variance than to ignore it. Linguistic resources (corpora and annotations) could be used to help the field characterize the unexplained variance both qualitatively and qualitatively, and to frame a research program to deal with uncertainty more effectively, assuming that it isn't possible to eliminate uncertainty, or even reduce it by much (enough that the remaining unexplained variance can be ignored).

Bio: Kenneth Church has worked on many topics in computational linguistics including: web search,language modeling, text analysis, spelling correction, word-sense disambiguation, terminology,translation, lexicography, compression, speech (recognition, synthesis & diarization), OCR, as well as applications that go well beyond computational linguistics such as revenue assurance and virtual integration (using screen scraping and web crawling to integrate systems that traditionally don't talk together as well as they could such as billing and customer care). He enjoys working with large corpora such as the Associated Press newswire (1 million words per week) and even larger datasets such as telephone call detail (1-10 billion records per month) and web logs. He earned his undergraduate and graduate degrees from MIT, and has worked at AT&T, Microsoft, Hopkins, IBM and Baidu. He was the president of ACL in 2012, and SIGDAT (the group that organizes EMNLP) from 1993 until 2011. He became an AT&T Fellow in 2001 and ACL Fellow in 2015.

12:15 – 1:45 – Lunch

1:45 – 2:15 – Denise Dipersio, Linguistic Data Consortium

"Development and Distribution of Clinical Speech and Language Datasets"

Abstract: As those involved in corpus development know, creating a data set requires attention to matters beyond the research question. These include regulatory issues that can control how, from whom or from where data can be collected, how it is displayed and how it can be used. For language resources in the clinical domain, such concerns are magnified to the extent a data set may contain individuals' personal health information, the disclosure of which is tightly controlled in the United States and elsewhere. Nevertheless, data sets for clinical uses exist and more are developed every day. This talk addresses the non-research aspects of language resource development and distribution in the clinical domain, including privacy, permissions and institutional review board/ethics board reviews. Available resources are also reviewed and different access policies compared.

Bio: Denise DiPersio is Associate Director at LDC where her principal responsibilities include the overall management of the External Relations Group in the areas of intellectual property rights, licensing, regulatory matters, distribution, publications, membership and communications.

2:15 – 2:45 – Mark Liberman, University of Pennsylvania:

"Clinical Applications of Human Language Technology"

Abstract: We infer a lot from the way someone talks: personal characteristics like age, gender, background, personality; contextual characteristics like mood and attitude towards the interaction; physiological characteristics like fatigue or intoxication. Many clinical diagnostic categories have symptoms that are partly or entirely manifest in spoken interaction: autism spectrum disorder, neurodegenerative disorders, schizophrenia, and so on. The development of modern speech and language technology makes it possible to create automated methods for diagnostic screening or monitoring. More important is the fact that these diagnostic categories are phenotypically diverse, representing (sometimes apparently discontinuous) regions of complex multidimensional behavioral spaces. We can hope that automated analysis of large relevant datasets will allow us to do better science, and learn what the true latent dimensions of those behavioral spaces are. And most important of all, we can hope for convenient, inexpensive, and psychometrically reliable ways to estimate the efficacy of treatments.

Bio: Mark Liberman received a PhD from MIT in 1975, and worked at Bell Labs Research from 1975 to 1990. Since 1990 he has been a faculty member at the University of Pennsylvania, with appointments in Linguistics and in Computer and Information Science. He founded the Linguistic Data Consortium in 1992.

2:45 – 3:15 – Hongwei Ding, Shanghai Jiao Tong University

"Construction of Linguistic Resources for Mental Disorders -- Interdisciplinary Research in Linguistics, Cognitive Neuroscience and Artificial Intelligence"

Abstract: This project is sponsored by the major program of national social science foundation of China (2018-2023). As linguists, we collaborate intensely with experts from clinical medicine, computer science and technology, speech pathology and neuroscience, psychology, and other disciplines. Our main task is to construct a speech database of mentally disordered people with normal control groups. These speech corpora will be supplemented with some video records and neurobehavioral data from both patients and control groups. Based on these speech-related databases, we will further explore the brain processing mechanism of speech and language in the disabled population and provide scientific guidance and basis for the relevant clinical research and treatment. At the same time, this project will try to combine audio, video, and neuroimaging features to build a multi-modal model to realize intelligent diagnosis to promote AI early screening of such mental diseases.

Bio: Hongwei DIng received her M.A. degree in Linguistics and Applied Linguistics from Shanghai Jiao Tong University, and her Dr. phil. and Dr.-Ing.habil. degrees from TU Dresden, Germany. She has gained extensive practical experiences in interdisciplinary fields, such as prosody and speech technology. Currently, she has focused her research interest on speech-language pathology, especially on patients with mental disorders and children with cochlear implant. She is an officer of the PAC (Permanent Advisory Committee) of SProSIG (Special Interest Group on Speech Prosody). She organized O-COCOSDA/CASLRE 2015 and Speech Prosody 2012 as program chair in Shanghai, and is working as workshop chair for 21th ACM International Conference on Multimodal Interaction 2019 in Suzhou.

3:15 – 3:30 – Coffee/Tea Break

3:30 – 4:00 – Rhoda Au, Neuropsychology Division, Framingham Heart Study

"Language Biomarkers of Brain Health"

Abstract: China has the largest aging population in the world and thus the largest at risk population for dementia. Currently there are no effective drug treatments for Alzheimer's disease (AD), the most common form of dementia. But technology is changing what we can do and how we can do it and digital voice offers an intriguing solution for the potential of detecting AD and other brain related disorders far earlier in its course, when potential disease modifying interventions may be more effective. The more transformative opportunity, however, is the prospect of digital voice indices serving as prognostic digital biomarkers, allowing for interventions that may drastically alter the disease trajectory so dramatically, it may be possible to prevent the disease altogether. But to enable these advancements in clinical research requires a paradigm shift from reliance on precedent based, hypothesis testing methodologies to embracing precedent-setting, hypothesis generating approaches. The long-term significance of this effort could would also lead to a conversion from the primary medical intervention model of "personalized medicine" to a more comprehensive focus on "personalized brain health".

Bio: Rhoda Au is Professor of Anatomy & Neurobiology, Neurology and Epidemiology at Boston University Schools of Medicine and Public. She also currently serves as Director of Neuropsychology at the Framingham Heart Study, where she has been involved in research related to cognitive aging and dementia since 1990. Most recently, she has been exploring the potential of cognitive digital biomarkers and how "big data" analytics can better inform our understanding of Alzheimer's disease pathways and treatment. Beyond Framingham, Dr. Au is focused on building multi-sector ecosystems to enable solutions for chronic disease prevention generally and optimizing brain health specifically and to move the primary focus of health technologies from precision medicine to a broader emphasis on precision health.

4:00 – 4:30 – Elif Eyigoz, IBM Thomas J. Watson Research Center

"Three applications of NLP: Drug intoxication, Schizophrenia, and Alzheimer's Disease"

Abstract: Automated analysis of language for accessing markers of mental conditions has gained immense interest in the last decade. This talk presents three studies on use of syntactic, semantic and acoustic analyses in this domain. For semantic analysis, we present our work on metaphor identification and sentiment-analysis for detection or prediction of schizophrenia. For acoustic analysis, we present our work on using phonological analysis for identification of subjects under the influence of two different drugs: Oxytocin (OT) and MDMA. For syntactic analysis, we present our work on using subtree pattern analysis for predicting severity of cognitive impairments, in particular Alzheimer's Disease.

Bio: Elif Eyigoz joined IBM research in 2016. Her research involves applications of natural language processing in accessing markers for cognitive impairments and psychiatric conditions. She got an MA in Linguistics at UCLA, an MS and a PhD in Computer Science at the University of Rochester. Previously, she got a BA in Philosophy and an MA in Cognitive Science at Bogazici University in Istanbul. During her studies, she worked on statistical machine translation models for morphologically rich languages, computational lexicography and syntax of languages with free word order. Before joining the research team at IBM, she worked as an NLP algorithms developer at Watson Labs at IBM.

4:30 – 5:00 – Tan Lee, The Chinese University of Hong Kong

"Deep learning approach to automatic detection of language impairment in spontaneous speech"

Abstract: Language impairment manifested in verbal communication is widely recognized as a useful biomarker for acquired and developmental disorders related to brain functions, e.g., aphasia, Parkinson's disease, and cognitive decline. In this talk, we present a deep learning approach to automatic detection and assessment of language impairment. A deep neural network based automatic speech recognition system is used to facilitate multi-scale acoustic and linguistic analysis of natural speech. Phone posteriorgrams, word embedding features, and supra-segmental duration extracted from the ASR output are found to be effective in discriminating impaired speech from unimpaired one. While our current work is carried out mainly with narrative speech from Cantonese-speaking people with aphasia, the proposed approach is generally applicable to other types of language disorders and other languages.

Bio: Tan Lee is an Associate Professor and the Director of the DSP and Speech Technology Laboratory at the Department of Electronic Engineering, the Chinese University of Hong Kong (CUHK). He has been working on speech and language related research since early 90s. In recent years, Tan Lee has been collaborating closely with medical doctors, and speech and hearing professionals, to apply advanced signal processing methods in dealing with human communication disorder problems. Examples of recent work include automatic assessment of voice disorder, speech disorder and language impairment, analysis of child and elderly speech and speech under pressure, unsupervised speech modeling for low-resource languages, and audio classification. Tan Lee is a member of IEEE and a member of the International Speech Communication Association (ISCA). He is an Associate Editor of the IEEE/ACM Transactions on Audio, Speech and Language Processing, and an editorial board member of the EURASIP Journal on Advances in Signal Processing. He served as an Area Chair for INTERSPEECH 2014, 2016 and 2018, and the General Co-Chair of the 2018 International Symposium on Chinese Spoken Language Processing. Tan Lee is currently the Vice Chair of ISCA Special Interest Group on Chinese Spoken Language Processing.

5:00 – 5:30 – Xiaowen Wang, Natalia Kiyueva, Emanuele Chersoni, and Chu-Ren

"The Curing Terms: From Medical Terminology to Causal Relations"

Abstract: Ongoing research combining medical databases and language resources for discovery of causal relations will be reported in this paper. More specifically, we aim to learn new information from observational data. Focussing on the domain of male infertility, we first extract terminology list from a corpus of medical research articles and enriched it with information from ontological resources. We take OHDSI (Observational Health Data Sciences and Informatics) Common Data Model as source data for causal relations, supplemented by WordNet and other relata, as well as syntactic patterns to extract all possible causal relations (including 'causes', 'due to', 'induced by', 'etiology of', etc.). Analysis of data from our pilot study confirms that this approach could lead to potential discovery of new causal relations. In addition, we also found that causal relations reported in observational data as well as medical papers are often descriptive causal relations and not logical direct causal relations. Such information can be enriched by combination with qualia eventive knowledge.

Bios: Xiaowen, Wang is an Associate Professor of applied linguistics in School of English and Education & Center for Institutional Discourse Studies, Guangdong University of Foreign Studies. She is currently a doctoral candidate in School of Humanities of the Hong Kong Polytechnic University under the supervision of Prof. Huang Chu-ren. Her research interests include corpus linguistics, English for medical purposes, ontology and discourse analysis.

Natalia Klyueva obtained her Ph.D. from Charles University in Prague in the field of computational linguistics (statistical machine translation). Currently working as a postdoctoral fellow at The Hong Kong Polytechnic University on various projects (medical texts, emotion detection, dialect identification) under the supervision of professor Chu-Ren Huang. The research interests are domain-independent and are focused mainly on technical aspects of text processing: statistical analysis, data manipulation as well as applying machine learning methods to various sorts of data.

Emmanuele Chersoni is currently a postdoctoral researcher in the Department of Chinese and Bilingual Studies of the Hong Kong Polytechnic University. He obtained a PhD in Theoretical and Computational Linguistics from Aix-Marseille University, working under the supervision of Prof. Philippe Blache (Aix-Marseille University) and Prof. Alessandro Lenci (University of Pisa). His main research interests include computational psycholinguistics, distributional semantic models, thematic fit modeling, automatic discovery of semantic relations and sentence processing.

Chu-Ren Huang (PhD, *Cornell* 1987; DHC, *Aix-Marseille* 2013) is a chair professor at the Hong Kong Polytechnic University. He has published 25 book or edited volumes, more than 25 online or licensable language resources, over 190 journal articles or book chapters, and over 450 refereed conference papers. His recent and upcoming books include *A Reference Grammar of Chinese, Computational Processing of the Chinese Language,* and *Cambridge Handbook of Chinese Linguistics (Cambridge); Mandarin Chinese Words and Parts of Speech: A corpus-based study,* and *Routledge Handbook in Chinese Applied Linguistics (Routledge); Digital Humanties: Bridging the Divide (Springer),* and *Generative Lexicon Studies in Chinese (Commercial Press).* He is Editor in Chief of the journal *Lingua Sinica* and the book series *SNLP (Cambridge), SEAL, THIA (Springer)* and *FiCL (PKU Press/Springer)*.