



# IALP 2019

## International Conference on Asian Language Processing

亚洲国际语言会议

---

### Organizers:

East China Normal University (ECNU)

Chinese and Oriental Languages Information Processing Society (COLIPS)

### Technical Sponsor:

IEEE Singapore SMC Chapter

15 November-17November, 2019

Shanghai, China

Shanghai Oriental Land Hotel 东方绿洲酒店



IALP 2019

## Overview of Conference Program

Date	Time	Event	Venue
Nov. 14	14:00-20:00	Onsite Registration	Shanghai Oriental Land Hotel (东方绿舟酒店)
Nov. 15	08:30-08:40	Opening	Sunshine Hall (阳光会堂)
	08:40-09:30	Keynote 1	Sunshine Hall
	Tea Break and Photo		
	10:10-11:25	Session 1 (AS, 5)	Sunshine Hall
	Short Tea Break		
	11:35-12:35	Session 2 (DD, 4)	Sunshine Hall
	Lunch		
	14:00-15:15	Session 3 (IE, 5)	Sunshine Hall
	Tea Break		
	15:30-16:45	Session 4 (LI, 5), Poster 1, Exhibition	Sunshine Hall, Meeting Room
	18:00-20:00	Banquet	Sunshine Hall
Nov. 16	08:30-09:20	Keynote 2	Meeting Room
	Tea Break		
	09:40-10:55	Session 5 (MT, 5)	Meeting Room
	Short Tea Break		
	11:10-12:25	Session 6 (SY, 5)	Meeting Room
	Lunch		
	14:00-15:00	Industry (3), Poster 2	Meeting Room
	Tea Break		
	15:15-16:30	Session 7(SE, 5)	Meeting Room
	Tea Break		
	16:45-18:00	Session 8 (LR, 5)	Meeting Room
	18:30-20:00	Dinner	
Nov. 17	08:30-09:45	Session 9 (QA, 5), Poster 3	Meeting Room
	Tea Break		
	10:00-11:15	Session 10 (AP, 5)	Meeting Room
	Tea Break		
	11:30-12:15	Best Paper, Next conference, Close	Meeting Room
	Lunch		
	14:00-	City Tour	

## Conference Program

November 14, 2019		
Time	Event	Venue
14:00-20:00	Onsite Registration	Shanghai Oriental Land Hotel
November 15, 2019		
08:30-08:40	Opening	Sunshine Hall
08:40-09:30	Keynote 1 (Xuanjing Huang)	Sunshine Hall
Tea Break (Photo)		
10:10-11:25	Session 1 (AS, 5)	Sunshine Hall
AS1	Acoustic Scene Classification Using Deep Convolutional Neural <i>Min Ye, Hong Zhong , Xiao Song , Shilei Huang, Gang Cheng</i>	
AS2	Acoustic Cues for Sensory Modality and Referential Vagueness of Personal Pronouns: Me or Not? You or Not? <i>Aibao Zhou, Yanbing Hu, Xiaoyong Lu, Tao Pan</i>	
AS3	Integrating Knowledge in End-to-End Automatic Speech Recognition for Mandarin- English Code-Switching <i>Chia-Yu Li, Ngoc Thang Vu</i>	
AS4	A Study on the Robustness of Pitch Range Estimation from Brief Speech <i>Wenjie Peng, Kaiqi Fu, Wei Zhang, Yanlu Xie, Jinsong Zhang</i>	
AS5	Improving Mandarin Prosody Boundary Detection by Using Phonetic Information and Deep LSTM Model <i>Ju Lin, Zhuanzhuan Ji, Wenwei Dong, Yanlu Xie, Jinsong Zhang</i>	
Short Tea Break		
11:35-12:35	Session 2 (DD, 4)	Sunshine Hall
DD1	Effective Data Augmentation Approaches to End-to-End Task-Oriented Dialogue <i>Jun Quan, Deyi Xiong</i>	
DD2	A Multi-stage Strategy for Chinese Discourse Tree Construction <i>Tishuang Wang, Peifeng Li, Qiaoming Zhu</i>	
DD3	A Systematic Investigation of Neural Models for Chinese Implicit Discourse Relationship Recognition <i>Dejian Li, Man Lan, Yuanbin Wu</i>	
DD4	Neural Machine Translation Strategies for Generating Honorific-style Korean <i>Lijie Wang, Mei Tu, Mengxia Zhai, Huadong Wang, Song Liu, Sang Ha Kim</i>	
Lunch		
14:00-15:15	Session 3 (IE, 5)	Sunshine Hall
IE1	Effect of Preprocessing for Distributed Representations: Case Study of Japanese Radiology Reports <i>Taro Tada, Kazuhide Yamamoto</i>	
IE2	Acquisition of Knowledge with Time Information from Twitter <i>Kohei Yamamoto, Kazutaka Shimada</i>	

IE3	Automatic Extraction and Quantitative Evaluation of the Character Relationship Networks from Children's Literature works <i>Kun Ma, Lijiao Yang</i>	
IE4	Employing Gated Attention and Multi-similarities to Resolve Document-level Chinese Event Coreference <i>Haoyi Cheng, Peifeng Li, Qiaoming Zhu</i>	
IE5	Multiple-source Entity Linking with Incomplete Sources <i>Qinguo Liu, Shui Liu, Lemao Liu, Bo Xiao</i>	
Tea Break		
15:30-16:45	Session 4 (LI, 5)	Sunshine Hall
LI1	Classified Description and Application of Chinese Constitutive Role <i>Mengxiang Wang, Cuiyan Ma</i>	
LI2	Exploring Letter's Differences between Partial Indonesian <i>Nankai Lin, Sihui Fu, Jiawen Huang, Shengyi Jiang</i>	
LI3	Comprehension correlates of the occurrence and deletion of “de” in Mandarin “N1 (de) N2” structures <i>Junyuan Zhao, Junru Wu</i>	
LI4	Effects of English Capitals On Reading Performance of Chinese Learners: Evidence from Eye Tracking <i>Yang Wei, Fu Xinyu</i>	
LI5	On the Etymology of he ‘river’ in Chinese <i>Huibin Zhuang, Zhanting Bu</i>	
15:30-16:45	Poster 1, Exhibition	Meeting Room
AP1 Poster	Quantifying the Use of English Words in Urdu News-Stories <i>Mehtab Alam Syed, Arif Ur Rahman, Muzammil Khan</i>	
AP2 Poster	Japanese Particle Error Correction employing Classification Model <i>Youichiro Ogawa, Kazuhide Yamamoto</i>	
AP3 Poster	Research on Chinese Text Error Correction Based on Sequence Model <i>Jianyong Duan Yang Yuan, Hao Wang, Xiaopeng Wei, Zheng Tan</i>	
AP4 Poster	Are Scoring Feedback of CAPT Systems Helpful for Pronunciation Correction? --An Exception of Mandarin Nasal Finals <i>Rui Cai, Wei Wei, Jinsong Zhang</i>	
AP5 Poster	Language Detection in Sinhala-English Code-mixed Data <i>Ian Smith, Uthayasanker Thayasivam</i>	
AP6 Poster	An experimental Tagalog Finite State Automata spellchecker with Levenshtein edit-distance feature <i>Joseph Marvin R. Imperial Czeritonnice Gail V. Ya-On, Jennifer C. Ureta</i>	
AP7 Poster	Chinese Spelling Check based on Sequence Labeling <i>Zijia Han, Zonglin Liu, Qiansheng Wang, Guohong Fu</i>	
AP8 Poster	Research on Tibetan Text Classification Method Based on Neural Network <i>Zhensong Li, Jie Zhu, Zhixiang Luo, Saihu Liu</i>	
AP9 Poster	Developing a machine learning-based grade level classifier for Filipino children's literature	

	<i>Joseph Marvin Imperial, Rachel Edita Roxas, Erica Mae Campos, Ani Rosa Almario</i>	
AP10 Poster	Development of a Filipino Speaker Diarization in Meeting Room Conversations <i>Angelica H. De La Cruz, Rodolfo C. Raga Jr</i>	
AP11 Poster	Construction of Quantitative Index System of Vocabulary Difficulty in Chinese Grade Reading <i>Huiping Wang,Lijiao Yang, Huimin Xiao</i>	
AP12 Poster	An Measurement Method of Ancient Poetry Difficulty for Adaptive Testing <i>Huiping Wang, Bihua Wang</i>	
AP13 Poster	Prosodic Realization of Focus in Changchun Mandarin and Nanjing Mandarin <i>Ying Chen, Jiajing Zhang, Bingying Ye, Chenfang Zhou</i>	
AP14 Poster	Celebrity Profiling from Twitter Data <i>Kumar Gourav Das, Braja Gopal Patra, Sudip Kumar Naskar</i>	
18:00-20:00	Banquet	Sunshine Hall
November 16, 2019		
08:30-09:20	Keynote 2 (Luo Si)	Meeting Room
Tea Break		
09:40-10:55	Session 5 (MT, 5)	Meeting Room
MT1	Character Decomposition for Japanese-Chinese Character-Level Neural Machine Translation <i>Jinyi Zhang, Tadahiro Matsumoto</i>	
MT2	Fusion of Image-text attention for Transformer-based Multimodal Machine Translation <i>Junteng Ma, Shihao Qin, Lan Su, Xia Li, Lixian Xiao</i>	
MT3	Statistical Machine Learning for Transliteration: Transliterating names between Sinhala, Tamil and English <i>H.S. Priyadarshani, M.D.W. Rajapaksha, M.M.S.P. Ranasinghe, K. Sarveswaran, G.V. Dias</i>	
MT4	Confidence Modeling for Neural Machine Translation <i>Taichi Aida, Kazuhide Yamamoto</i>	
MT5	Syntax-aware Transformer Encoder for Neural Machine Translation <i>Sufeng Duan, Hai Zhao, Junru Zhou, Rui Wang</i>	
Short Tea Break		
11:10-12:25	Session 6 (SY, 5)	Meeting Room
SY1	What affects the difficulty of Chinese syntax? <i>Yueming Du, Lijiao Yang</i>	
SY2	A Chinese word segment model for energy literature based on Neural Networks with Electricity User Dictionary <i>Bochuan Song, Bo Chai, Qiang Zhang, Quanye Jia</i>	
SY3	A Study on Syntactic Complexity and Text Readability of ASEAN English News <i>Yusha Zhang, Nankai Lin, Shengyi Jiang</i>	
SY4	Tibetan word segmentation method based on CNN-BiLSTM-CRF model <i>Lili Wang, Hongwu Yang, Xiaotian Xing, Yajing Yan</i>	

SY5	Converting an Indonesian Constituency Treebank to the Penn Treebank Format <i>Jessica Naraiswari Arwidarasti, Ika Alfina, Adila Alfa Krisnadhi</i>	
Lunch		
14:00-15:00	Industry (3)	Meeting Room
Industry		
14:00-15:00	Poster 2	Meeting Room
QA1 Poster	Duplicate Question Detection based on Neural Networks and Multi-head Attention <i>Heng Zhang,Liangyu Chen</i>	
QA2 Poster	Automated Prediction of Item Difficulty in Reading Comprehension Using Long Short-Term Memory <i>Li-Huai Lin, Tao-Hsing Chang, Fu-Yuan Hsu</i>	
QA3 Poster	Improving Question Classification with Hybrid Networks <i>Yichao Cao, Miao Li, Tao Feng, Ruijing Wang, Yue Wu</i>	
QA4 Poster	Separate Answer Decoding for Multi-class Question Generation <i>Kaili Wu, Yu Hong, Mengmeng Zhu, Hongxuan Tang, Min Zhang</i>	
QA5 Poster	How to Answer Comparison Questions <i>Hongxuan Tang, Yu Hong, Xin Chen, Kaili Wu, Min Zhang</i>	
QA6 Poster	Coarse-to-Fine Document Ranking for Multi-Document Reading Comprehension with Answer-Completion <i>Hongyu Liu, Shumin Shi, Heyan Huang</i>	
SE1 Poster	Aspect-based Opinion Mining for Code-Mixed Restaurant Reviews in Indonesia <i>Andi Suciati, Indra Budi</i>	
SE2 Poster	CIEA: A Corpus for Chinese Implicit Emotion Analysis <i>Dawei Li, Jin Wang, Xuejie Zhang</i>	
SE3 Poster	An Enhancement of Malay Social Media Text Normalization for Lexicon-Based Sentiment Analysis <i>Muhammad Fakhrur Razi Abu Bakar, Norisma Idris, Liyana Shuib</i>	
SE4 Poster	Sinhala and Tamil Speech Intent Identification From English Phoneme Based ASR <i>Yohan Karunanayake, Uthayasanker Thayasivam, Surangika Ranathunga</i>	
SE5 Poster	Exploring Characteristics of Word Co-occurrence Network in Translated Chinese <i>Jianyu Zheng, Kun Ma, Xuemei Tang, Shichen Liang</i>	
SE6 Poster	Improving Japanese-English Bilingual Mapping of Word Embeddings based on Language Specificity <i>Yuting Song, Biligsaikhan Batjargal, Akira Maeda</i>	
SE7 Poster	Exploring Context's Diversity to Improve Neural Language Model <i>Yanchun Zhang, Xingyuan Chen, Peng Jin, Yajun Du</i>	
SE8 Poster	A General Procedure for Improving Language Models in Low-Resource Speech Recognition <i>Qian Liu, Wei-Qiang Zhang, Jia Liu, Yao Liu</i>	
Tea Break		
15:15-16:30	Session 7(SE, 5)	Meeting Room

SE1	Combination of Semantic Relatedness with Supervised Method for Word Sense Disambiguation <i>Qiaoli Zhou, Yuguang Meng</i>	
SE2	Levergeing Character embedding to Chinese Textual Entailment Recognition Based on Attention <i>Pengcheng Liu, Lingling Mu, Hongying Zan</i>	
SE3	Automatic Recognition of Allusions in Tang Poetry Based on BERT <i>Xuemei Tang , Shichen Liang, Jianyu Zheng, Renfen Hu, Zhiying Liu</i>	
SE4	Diachronic Synonymy and Polysemy: Exploring Dynamic Relation Between Forms and Meanings of Words Based on Word Embeddings <i>Shichen Liang, Jianyu Zheng, Xuemei Tang, Renfen Hu, Zhiying Liu</i>	
SE5	Carrier Sentence Selection with Word and Context Embeddings <i>Chak Yan Yeung, John Lee, Benjamin Tsou</i>	
Tea Break		
16:45-18:00	Session 8 (LR, 5)	Meeting Room
LR1	A New Method of Tonal Determination for Chinese Dialects <i>Yan Li, Zhiyi Wu</i>	
LR2	Extremely Low Resource Text simplification with Pre-trained Transformer <i>Takumi Maruyama, Kazuhide Yamamoto</i>	
LR3	Automatic Meta-evaluation of Low-Resource Machine Translation Evaluation Metrics <i>Junting Yu, Wuying Liu , Hongye He , Lin Wang</i>	
LR4	The Initial Research of Mongolian Literary Corpus-Take the Text of Da.Nachugdorji's Work for Instance <i>YinhuaHai</i>	
LR5	Statistical Analysis of Syllable Duration of Uyghur Language <i>Askar Hamdulla, Guzalnur Dilmurat, Gulnur Arkin, Mijit Ablimit</i>	
18:30-20:00	Dinner	
November 17, 2019		
08:30-09:45	Session 9 (QA, 5)	Meeting Room
QA1	Examination-Style Reading Comprehension with Neural augmented Retrieval <i>Yiqing Zhang, HaiZhao, ZhuoshengZhang</i>	
QA2	Automatic answer ranking based on sememe vector in KBQA <i>Yadi Li, Lingling Mu, Hao Li, Hongying Zan</i>	
QA3	Towards Robust Neural Machine Reading Comprehension via Question Paraphrases <i>Ying Li, Hongyu Li, Jing Liu</i>	
QA4	Using WHY-type Question-Answer Pairs to Improve Implicit Causal Relation Recognition <i>Huibin Ruan, Yu Hong, Yu Sun, Yang Xu, Min Zhang</i>	
QA5	Ranking Like Human: Global-View Matching via Reinforcement Learning for Answer Selection <i>Yingxue Zhang, Ping Jian, Ruiying Geng, Yuansheng Song, Fandong Meng</i>	
08:30-09:45	Poster 3	Meeting Room

IE1 Poster	A Machine Learning Model for Dating of Ancient Chinese Texts <i>Xuejin Yu, Wei Huangfu</i>	
IE2 Poster	Using Mention Segmentation to Improve Event Detection with Multi-head Attention <i>Jiali Chen, Yu Hong, Jingli Zhang, Jianmin Yao</i>	
IE3 Poster	An End-to-End Model Based on TDNN-BiGRU for Keyword Spotting <i>Shuzhou Chai, Zhenyu Yang, Changsheng Lv, Wei-Qiang Zhang</i>	
IE4 Poster	Research on New Event Detection Methods for Mongolian News <i>Shijie Wang, Feilong Bao, Guanglai Gao</i>	
MT1 Poster	Cross Language Information Retrieval Using Parallel Corpus with Bilingual Mapping Method <i>Rinaldi Andrian Rahmanda, Mirna Adriani, Dipta Tanaya</i>	
MT2 Poster	Phrase-Based Tibetan-Chinese Statistical Machine Translation <i>YONG Cuo, Xiaodong SHI, NYIMA Tashi, Yidong CHEN</i>	
AS1 Poster	Design and Implementation of Burmese Speech Synthesis System Based on HMM-DNN <i>Mengyuan Liu, Jian Yang</i>	
AS2 Poster	Speech Command Classification System for Sinhala Language based on Automatic Speech Recognition <i>Thilini Dinushika, Lakshika Kavmini, Pamoda Abeyawardhana, Uthayasanker Thayasivam, Sanath Jayasena</i>	
AS3 Poster	Articulatory Features Based TDNN Model for Spoken Language Recognition <i>Jiawei Yu, Minghao Guo, Yanlu Xie, Jinsong Zhang</i>	
AS4 Poster	Improved DNN-HMM English Acoustic Model Specially For Phonotactic Language Recognition <i>Wei-Wei Liu, Guo-Chun Li, Cun-Xue Zhang, Hai-Feng Yan, Jing He, Ying-Xin Gan, Yan-Miao Song, Jian-Hua Zhou, Jian-Zhong Liu, Ying Yin, Ya-Nan Li*, Yu-Bin Huang, Ting Ruan, Wei Liu, Rui-Li Du, Hua-ying Bai, Wei Li, Sheng-Ge Zhang</i>	
AS5 Poster	Comparison of Prosodic Encoding of Focus in Chinese Formal and Informal Speech <i>Xin Zhong, Hui Feng</i>	
AS6 Poster	A Comparative Analysis of Acoustic Characteristics between Kazak & Uyghur Mandarin Learners and Standard Mandarin Speakers <i>Gulnur Arkin, Gvljan Alijan, Askar Hamdulla, Mijit ablimit</i>	
Tea Break		
10:00-11:15	Session 10 (AP, 5)	Meeting Room
AP1	Japanese grammatical simplification with simplified corpus <i>Yumeto Inaoka, Kazuhide Yamamoto</i>	
AP2	Learning Deep Matching-Aware Network for Text Recommendation using Clickthrough Data <i>Haonan Liu, Nankai Lin, Zitao Chen, Ke Li and Shengyi Jiang</i>	
AP3	Correlational Neural Network Based Feature <i>Wenwei Dong, Yanlu Xie</i>	
AP4	Improving text simplification by corpus expansion with unsupervised learning <i>Akihiro Katsuta, Kazuhide Yamamoto</i>	



AP5	BERT with Enhanced Layer for Assistant Diagnosis Based on Chinese Obstetric EMRs <i>Kunli Zhang, Chuang Liu, Xuemin Duan, Lijuan Zhou, Yueshu Zhao, Hongying Zan</i>	
Tea Break		
11:30-12:00	Best Paper Award, Next Conference, Close	Meeting Room
Lunch		
14:00-	City Tour	

## Paper Information

November 15, 2019		
10:10-11:25	Session 1 (AS, 5)	Sunshine Hall
<b>Acoustic Scene Classification Using Deep Convolutional Neural</b> <i>Min Ye, Hong Zhong, Xiao Song, Shilei Huang, Gang Cheng</i> <p>We use deep convolutional neural network via transfer learning for Acoustic Scene Classification (ASC). For this purpose, a powerful and popular deep learning architecture — Residual Neural Network (Resnet) is adopted. Transfer learning is used to fine-tune the pre-trained Resnet model on the TUT Urban Acoustic Scenes 2018 dataset. Furthermore, the focal loss is used to improve overall performance. In order to reduce the chance of overfitting, data augmentation technique is applied based on mixup. Our best system has achieved an improvement of more than 10% in terms of class-wise accuracy with respect to the Detection and classification of acoustic scenes and events (DCASE) 2018 baseline system on the TUT Urban Acoustic Scenes 2018 dataset.</p>		
<b>Acoustic Cues for Sensory Modality and Referential Vagueness of Personal Pronouns: Me or Not? You or Not?</b> <i>Aibao Zhou, Yanbing Hu, Xiaoyong Lu, Tao Pan</i> <p>Personal pronouns are of great pragmatic significance, and among their numerous functions, referential vagueness is the key to the exploration of the self. In Experiment 1, the relationship of first-, second-, and third- personal pronouns with the self was discussed in acoustic condition, with gender consistency, pronoun type, and person type as independent variables. Experiment 2 records the frequency of the subjects on the SR button and the NSR button, and statistical analysis was performed on the SR button reaction. This study finds that the same pronouns show different self-cognitive processing under the different visual and acoustic stimuli, and the results support the dynamics of personal pronouns and possessive pronouns and selfrelationships.</p>		
<b>Integrating Knowledge in End-to-End Automatic Speech Recognition for Mandarin-English Code-Switching</b> <i>Chia-Yu Li, Ngoc Thang Vu</i> <p>Code-Switching (CS) is a common linguistic phenomenon in multilingual communities that consists of switching between languages while speaking. This paper presents our investigations on end-to-end speech recognition for Mandarin-English CS speech. We analyze different CS specific issues such as the properties mismatches between languages in a CS language pair, the unpredictable nature of switching points, and the data scarcity problem. We exploit and improve the state-of-the-art end-to-end system by merging nonlinguistic symbols, by integrating language identification using hierarchical softmax, by modeling subword units, by artificially lowering the speaking rate, and by augmenting data using speed perturbed technique and several monolingual datasets to improve the final performance not only on CS speech but also on monolingual benchmarks in order to making the system more applicable on real life settings. Finally, we explore the effect of different language model integration methods on the performance of the proposed model. Our experimental results reveal that all the proposed techniques improve the recognition performance. The best combined system improves the baseline system by up to</p>		

35% relatively in terms of mixed error rate and delivers acceptable performance on monolingual benchmarks.

#### **A Study on the Robustness of Pitch Range Estimation from Brief Speech**

*Wenjie Peng, Kaiqi Fu, Wei Zhang, Yanlu Xie, Jinsong Zhang*

Pitch range estimation from brief speech segments is important for many tasks like automatic speech recognition. To address this issue, previous studies have proposed to utilize deep-learning-based models to estimate pitch range with spectrum information as input [1-2]. They demonstrated it could still achieve reliable estimation results when speech segment is as brief as 300ms. In this work, we further investigate the robustness of this method. We take the following situation into account: 1) increasing the number of speakers for model training hugely; 2) secondlanguage( L2) speech data; 3) the influence of monosyllabic utterances with different tones. We conducted experiments accordingly. Experimental results showed that: 1) We further improved the accuracy of pitch range estimation after increasing the speakers for model training. 2) The estimation accuracy on the L2 learners is similar to that on the native speakers. 3) Different tonal information has an influence on the LSTM-based model, but this influence is limited compared to the baseline method. These results may contribute to speech systems that demanding pitch features.

#### **Improving Mandarin Prosody Boundary Detection by Using Phonetic Information and Deep LSTM Model**

*Ju Lin, Zhuanzhuan Ji, Wenwei Dong, Yanlu Xie, Jinsong Zhang*

Automatic prosodic boundary detection is useful for automatic speech processing, such as automatic speech recognition (ASR) and speech synthesis. In this paper, we propose two techniques to improve the boundary detection performance. First, in addition to prosody features (e.g, pitch, duration and energy), phonetic information (word/articulatory information) is integrated into the framework of prosodic boundary detection. We compared two forms of phonetic information: word form and articulatory form. Moreover, boundary detection can be regarded as a sequence labeling task. A deep Long Short-Term Memory (LSTM) is adopted for this task, which replaces the traditional Deep Neural Networks (DNN) model. The experimental results showed that the boundary detection performance can be improved by the additional phonetic information, with relative 5.9% (word form) and 9.8% (articulatory form) improvements respectively in contrast with the system that only used prosody features modeled. The articulatory information and prosody features with deep LSTM achieved the best result, with further performance enhancement from 76.35% to 77.85% (relative 6.3%) compared with that modeled by DNN.

**11:35-12:35**

**Session 2 (DD, 4)**

**Sunshine Hall**

#### **Effective Data Augmentation Approaches to End-to-End Task-Oriented Dialogue**

*Jun Quan, Deyi Xiong*

The training of task-oriented dialogue systems is often confronted with the lack of annotated data. In contrast to previous work which augments training data through expensive crowd-sourcing efforts, we propose four different automatic approaches to data augmentation at both the word and sentence level for end-to-end task-oriented dialogue and conduct an empirical study on their impact. Experimental results on the CamRest676 and KVRET datasets demonstrate that each of the four data augmentation approaches is able to obtain a significant improvement over a strong baseline in terms of Success F1

score and that the ensemble of the four approaches achieves the state-of-the-art results in the two datasets. In-depth analyses further confirm that our methods adequately increase the diversity of user utterances, which enables the end-to-end model to learn features robustly.

#### **A Multi-stage Strategy for Chinese Discourse Tree Construction**

*Tishuang Wang, Peifeng Li, Qiaoming Zhu*

Building discourse tree is crucial to improve the performance of discourse parsing. There are two issues in previous work on discourse tree construction, i.e., the error accumulation and the influence of connectives in transitionbased algorithms. To address above issues, this paper proposes a tensor-based neural network with the multi-stage strategy and connective deletion mechanism. Experimental results on both CDTB and RST-DT show that our model achieves the state-of-the-art performance.

#### **A Systematic Investigation of Neural Models for Chinese Implicit Discourse Relationship Recognition**

*Dejian Li, Man Lan, Yuanbin Wu*

Building discourse tree is crucial to improve the performance of discourse parsing. There are two issues in previous work on discourse tree construction, i.e., the error accumulation and the influence of connectives in transitionbased algorithms. To address above issues, this paper proposes a tensor-based neural network with the multi-stage strategy and connective deletion mechanism. Experimental results on both CDTB and RST-DT show that our model achieves the state-of-the-art performance.

#### **Neural Machine Translation Strategies for Generating Honorific-style Korean**

*Lijie Wang, Mei Tu, Mengxia Zhai, Huadong Wang, Song Liu, Sang Ha Kim*

Expression with honorifics is an important way of dressing up the language and showing politeness in Korean. For machine translation, generating honorifics is indispensable on the formal occasion when the target language is Korean. However, current Neural Machine Translation (NMT) models ignore generation of honorifics, which causes the limitation of the MT application on business occasion. In order to address the problem, this paper presents two strategies to improve Korean honorific generation ratio: 1) we introduce honorific fusion training (HFT) loss under the minimum risk training framework to guide the model to generate honorifics; 2) we introduce a data labeling (DL) method which tags the training corpus with distinctive labels without any modification to the model structure. Our experimental results show that the proposed two strategies can significantly improve the honorific generation ratio by 34.35% and 45.59%.

14:00-15:15

Session 3 (IE, 5)

Sunshine Hall

#### **Effect of Preprocessing for Distributed Representations: Case Study of Japanese Radiology Reports**

*Taro Tada, Kazuhide Yamamoto*

A radiology report is a medical document based on an examination image in a hospital. However, the preparation of this report is a burden on busy physicians. To support them, a retrieval system of past documents to prepare radiology reports is required. In recent years, distributed representation has been used in various NLP tasks and its usefulness has been demonstrated. However, there is not much research about Japanese medical documents that use distributed representations. In this study, we investigate preprocessing on a retrieval system with a distributed representation of the radiology report, as a first step. As a result, we confirmed that in word segmentation using Morphological analyzer and dictionaries,

medical terms in radiology reports are not handled as long nouns, but are more effective as shorter nouns like subwords. We also confirmed that text segmentation by SentencePiece to obtain sentence distributed representation reflects more sentence characteristics. Furthermore, by removing some phrases from the radiology report based on frequency, we were able to reflect the characteristics of the document and avoid unnecessary high similarity between documents. It was confirmed that preprocessing was effective in this task.

#### **Acquisition of Knowledge with Time Information from Twitter**

***Kohei Yamamoto, Kazutaka Shimada***

In this paper, we propose a knowledge acquisition method for non-task-oriented dialogue systems. Such dialogue systems need a wide variety of knowledge for generating appropriate and sophisticated responses. However, constructing such knowledge is costly. To solve this problem, we focus on a relation about each tweet and the posted time. First, we extract event words, such as verbs, from tweets. Second, we generate frequency distribution for five different time divisions: e.g., a monthly basis. Then, we remove burst words on the basis of variance for obtaining refined distributions. We checked high ranked words in each time division. As a result, we obtained not only common sense things such as "sleep" in night but also interesting activities such as "recruit" in April and May (April is the beginning of the recruitment process for the new year in Japan.) and "raise the spirits/plow into" around 9 AM for inspiring oneself at the beginning of his/her work of the day. In addition, the knowledge that our method extracts probably contributes to not only dialogue systems but also text mining and behavior analysis of data on social media and so on.

#### **Automatic Extraction and Quantitative Evaluation of the Character Relationship Networks from Children's Literature works**

***Kun Ma, Lijiao Yang***

To automate the graded reading task, we urgently need to extract and calculate the important index of the complexity of the relationship between the characters affecting the plot complexity of narrative literature. In order to realize this purpose, this paper describes a computational method for automatic analysis of the virtual social network from children's literature works. We selected the required bibliography for primary school students recommended by the Ministry of Education, then automatically extract the characters of the novel by CRF, and constructs the character network based on the co-occurrence relationship. The statistical analysis method of complex network provides a quantitative basis for distinguishing the complexity of characters' relationships in different texts. The results show that the structural characteristics of character interaction networks are similar to those of small world networks, and the selected network measurement indexes are significantly related to the complexity of text characters. Finally, we achieved effectively evaluating and predicting the complexity of the social networks from more extensive literature works some classical regression model based on machine learning.

#### **Employing Gated Attention and Multi-similarities to Resolve Document-level Chinese Event Coreference**

***Haoyi Cheng, Peifeng Li, Qiaoming Zhu***

To automate the graded reading task, we urgently need to extract and calculate the important index of the complexity of the relationship between the characters affecting the plot complexity of narrative literature. In order to realize this purpose, this paper describes a computational method for automatic

analysis of the virtual social network from children's literature works. We selected the required bibliography for primary school students recommended by the Ministry of Education, then automatically extract the characters of the novel by CRF, and constructs the character network based on the co-occurrence relationship. The statistical analysis method of complex network provides a quantitative basis for distinguishing the complexity of characters' relationships in different texts. The results show that the structural characteristics of character interaction networks are similar to those of small world networks, and the selected network measurement indexes are significantly related to the complexity of text characters. Finally, we achieved effectively evaluating and predicting the complexity of the social networks from more extensive literature works some classical regression model based on machine learning.

#### **Multiple-source Entity Linking with Incomplete Sources**

*Qinguo Liu, Shui Liu, Lemao Liu, Bo Xiao*

This paper introduces a new entity linking task from a well-known online video application in industry, where both entities and mentions are represented by multiple sources but some of them may be missing. To address the issue of incomplete sources, it proposes a novel neural approach to model the linking relationship between a pair of an entity and a mention. To verify the proposed approach to this task, it further creates a large scale dataset including 70k examples. Experiments on this dataset empirically demonstrate that the proposed approach is effective over a baseline and particularly it is robust to the missing sources in some extent.

15:30-16:45

Session 4 (LI, 5)

Sunshine Hall

#### **Classified Description and Application of Chinese Constitutive Role**

*Mengxiang Wang, Cuiyan Ma*

Constitutive role is one of the 4 qualia roles, which expresses a kind of constitutive relationship between nouns. According to the original definition and description characteristics, this paper divides the constitutive roles into two categories: materials and components. At the same time, combined with the previous methods of extracting the role automatically, this paper optimizes the method of extracting the role automatically. Relying on auxiliary grammatical constructions, we extract noun-noun pairs from large-scale corpus to extract descriptive features of constitutive roles, and then classifies these descriptive knowledge by manual double-blind proofreading. Finally, the author discusses the application of Chinese constitutive roles in word-formational analysis, syntactic analysis and synonym discrimination.

#### **Exploring Letter's Differences between Partial Indonesian**

*Nankai Lin, Sihui Fu, Jiawen Huang, Shengyi Jiang*

Differences of letter usage are the most basic differences between different languages, which can reflect the most essential diversity. Many linguists study the letter differences between common languages, but seldom research those between non-common languages. This paper selects three representative languages from the Indonesian branch of the Austronesian language family, namely Malay, Indonesian and Filipino. To study the letter differences between these three languages and English, we concentrate on word length distribution, letter frequency distribution, commonly used letter pairs, commonly used letter trigrams, and ranked letter frequency distribution. The results show that great differences do exist between three Indonesianbranch languages and English, and the differences between Malay and Indonesian are the smallest.

**Comprehension correlates of the occurrence and deletion of “de” in Mandarin “N1 (de) N2” structures**

*Junyuan Zhao, Junru Wu*

Based on corpus materials and on-line semantic judgment surveys, this paper investigates the comprehension differences related to the occurrence and deletion of “de” in the Mandarin “N1 (de) N2” structure. By applying PCA and LME modellings on a set of semantic survey data, this study provides a multi-level database of semantic measurements for a set of Chinese “N1 (de) N2” structures as well as a quantitative analysis regarding the correlation between structure-level and constituent-level semantic features. The research shows that: (1) The “de”-occurring structure is more likely to be interpreted as indefinite than the “de”-deletion structure. (2) Animacy of N1 is positively related to the grammaticality of the “de”-occurring structure, while animacy of N1 is negatively related to the grammaticality of the “de”-deletion structure. The research findings provide evidence for prototype effects in the process of language comprehension. We propose that in natural comprehension, there is a high-animacy bias for N1 regarding the “de”-occurring structure; while a low animacy interpretation for N1 is more prototypical for the “de”-deletion structure. Accordingly, the “de”-occurring structure tends to be interpreted as a possessive, while the “de”-deletion structure is more likely to be interpreted as a modifier-head structure.

**Effects of English Capitals On Reading Performance of Chinese Learners: Evidence from Eye Tracking**

*Yang Wei, Fu Xinyu*

Native English speakers need more time to recognize capital letters in reading, yet the influence of capitals upon Chinese learners' reading performance is seldom studied. We conducted an eye tracker experiment to explore the cognitive features of Chinese learners in reading texts containing capital letters. Effect of English proficiency on capital letter reading is also studied. The results showed that capitals significantly increase the cognitive load in Chinese learners' reading process, complicate their cognitive processing, and lower their reading efficiency. The perception of capital letters of Chinese learners is found to be an isolated event and may influence the word superiority effect. English majors, who possess relatively stronger English logical thinking capability than non-English majors, face the same difficulty as the non-English majors do if no practice of capital letter reading have been done.

**On the Etymology of he ‘river’ in Chinese**

*Huibin Zhuang, Zhanting Bu*

In Chinese he 河 ‘river’ can be used as proper names (for the Yellow River) as well as a common word for rivers in North China. Based on linguistic data, ethnological evidence and historical documents, this paper argues against these leading hypotheses and proposes that he originated from the Old Yi language, entered Chinese through language contact and replaced shui which was from Old Qiang and later became the only common noun for river in North China.

15:30-16:45

Poster 1, Exhibition

Meeting Room

**Quantifying the Use of English Words in Urdu News-Stories**

*Mehtab Alam Syed, Arif Ur Rahman, Muzammil Khan*

The vocabulary of Urdu language is a mixture of many other languages including Farsi, Arabic and Sanskrit. Though, Urdu is the national language of Pakistan, English has the status of official language of Pakistan. The use of English words in spoken Urdu as well as documents written in Urdu is increasing with

the passage of time. The automatic detection of English words written using Urdu script in Urdu text is a complicated task. This may require the use of advanced machine/deep learning techniques. However, the lack of initial work for developing a fully automatic system makes it a more challenging task. The current paper presents the result of an initial work which may lead to the development of an approach which may detect any English word written Urdu text. First, an approach is developed to preserve Urdu stories from online sources in a normalized format. Second, a dictionary of English words transliterated into Urdu was developed. The results show that there can be different categories of words in Urdu text including transliterated words, words originating from English and words having exactly similar pronunciation but different meaning.

#### **Japanese Particle Error Correction employing Classification Model**

*Youichiro Ogawa, Kazuhide Yamamoto*

We present a grammatical error correction system for Japanese particles based on the classification method. We define a confusion set of the particles for detection of particle errors and prediction of the correct word. Our method can handle not only substitutions but also insertions and deletions. For building the training data, we used two datasets: a large amount of native language data and corrected learners' sentences. That is, we did not require a parallel corpus of learners. We show the results for Japanese particle error correction on the NAIST Goyo corpus, evaluated by the MaxMatch (M2) score. In addition, we analyze the effect of percentage changes in deletion labels while building the training data and analyze the prediction probability threshold at correction. Our best model achieved 46.4 F<sub>0.5</sub>.

#### **Research on Chinese Text Error Correction Based on Sequence Model**

*Jianyong Duan Yang Yuan, Hao Wang, Xiaopeng Wei, Zheng Tan*

When users input text, it will inevitably produce errors, and with the rapid development and popularization of smart devices, the situation becomes more and more serious. Therefore, text correction has become one of the important research directions in the field of natural language processing. As the grammatical error correction task, in this paper, the error correction process of Chinese text is regarded as the conversion process from wrong sentence to correct sentence. In order to adapt to this task, the (sequence-to-sequence) Seq2Seq model is introduced. The wrong sentence is used as the source sentence, and the correct sentence is used as the target sentence. Supervised training is carried out in units of characters and words. It can be used for correcting errors such as word of homophone, homotype, and near-sound, greatly reducing the artificial participation and expert support of feature extraction, improve model accuracy on specific errors. In order to solve the information loss caused by the conversion of long sequence to fixed length vector, the attention mechanism is introduced into the basic model. After adding the attention mechanism, the model's accuracy, recall rate and F1 value have been effectively improved.

#### **Are Scoring Feedback of CAPT Systems Helpful for Pronunciation Correction? --An Exception of Mandarin Nasal Finals**

*Rui Cai, Wei Wei, Jinsong Zhang*

The scoring feedback of Computer Assisted Pronunciation Training (CAPT) systems facilitate learner's instant awareness of their problems, easily lead to more practices. But whether it is enough to instruct the learners to understand how to correct their errors is still unknown. To see in depth, the impacts from CAPT technology on language learning, and to investigate learner's correction strategy after receiving



error warnings, this paper studies long term learning data of Chinese utterances by a number of CSL (Chinese as a Second Language) learners, with special efforts paid to the utterances of nasal Finals. The data resulted from a 3-week use of a CAPT APP, called "SAIT 汉语" for Chinese learning, by 10 learners with different mother tongues. Major findings include: 1) Improvements were seen with almost all kinds of phonemes, except nasal Finals; 2) Data analyses showed that the learners had tried to lengthen the nasal codas after they received error warnings, while Chinese native data shows a significant nasalization period before a short coda. These results suggest that the scoring feedback can be beneficial to pronunciation training in most cases, except for some special ones. For the sounds such as Chinese nasal Finals, more appropriate feedback method is desired.

#### **Language Detection in Sinhala-English Code-mixed Data**

***Ian Smith, Uthayasanker Thayasivam***

Language identification in text data has become a trending topic due to multiple language usage on the internet and it becomes a difficult task when it comes to bilingual and multilingual communication data processing. Accordingly, this study introduces a methodology to detect Sinhala and English words in code-mixed data and this is the first research done on such scenario at the time of this paper is written. In addition to that, the data set which is used for this research was newly built and published for similar research users. Even though there are well known models to identify Singlish Unicode characters which is a straightforward study; there are no proper language detection models to detect Sinhala words in a sentence which contains English words (code-mixed data). Therefore, this paper presents a language detection model with XGB classifier with 92.1% accuracy and a CRF model with a F1-score of 0.94 for sequence labeling.

#### **An experimental Tagalog Finite State Automata spellchecker with Levenshtein edit-distance feature**

***Joseph Marvin R. Imperial Czeritonnio Gail V. Ya-On, Jennifer C. Ureta***

In this paper, we present an experimental development of a spell checker for the Tagalog language using a set of word list with 300 random root words and three inflected forms as training data and a two-layered architecture of combined Deterministic Finite Automaton (DFA) with Levenshtein edit-distance. A DFA is used to process strings to identify if it belongs to a certain language via the binary result of accept or reject. The Levenshtein edit-distance of two strings is the number (k) of deletions, alterations, insertions between two sequences of characters. From the sample trained wordlist, results show that a value of 1 for the edit-distance (k) can be effective in spelling Tagalog sentences. Any value greater than 1 can cause suggestion of words even if the spelling of words is correct due to selective and prominent usage of certain characters in the Tagalog language like a, n, g, t, s, l.

#### **Chinese Spelling Check based on Sequence Labeling**

***Zijia Han, Zonglin Liu, Qiansheng Wang, Guohong Fu***

Chinese has been widely used by people from all over the world. Various Chinese spelling errors may occur inevitably from Chinese as Foreign Language (CFL) beginners. This paper presents a method for Chinese spelling check to detect and correct spelling errors in a sentence. Our approach is based on the sequence labeling model using the bi-direction LSTM network. We also utilize the Pinyin feature and Chinese strokes feature to improve performance. The evaluation on the SIGHAN-8 shows that our approach gets good performance on both detection and correction tasks.

### **Research on Tibetan Text Classification Method Based on Neural Network**

***Zhensong Li, Jie Zhu, Zhixiang Luo, Saihu Liu***

Text categorization is an important task in natural language processing, and it has a wide range of applications in real life. In this paper, two N-Gram feature models (MLP, FastText) and two sequential models (sepCNN, Bi-LSTM) are used to study the automatic classification for Tibetan text based on syllables and vocabulary. The experiment on Tibetan language data collected by China Tibet News Network shows that the classification accuracy is about 85%.

### **Developing a machine learning-based grade level classifier for Filipino children's literature**

***Joseph Marvin Imperial, Rachel Edita Roxas, Erica Mae Campos, Ani Rosa Almario***

Reading is an essential part of children's learning. Identifying the proper readability level of reading materials will ensure effective comprehension. We present our efforts to develop a baseline model for automatically identifying the readability of children's and young adult's books written in Filipino using machine learning algorithms. For this study, we processed 258 picture books published by Adarna House Inc. In contrast to old readability formulas relying on static attributes like number of words, sentences, syllables, etc., other textual features were explored. Count vectors, Term Frequency-Inverse Document Frequency (TF-IDF), n-grams, and character-level n-grams were extracted to train models using three major machine learning algorithms—Multinomial Naive-Bayes, Random Forest, and K-Nearest Neighbors. A combination of K-Nearest Neighbors and Random Forest via voting-based classification mechanism resulted with the best performing model with a high average training accuracy and validation accuracy of 0.822 and 0.74 respectively. Analysis of the top 10 most useful features for each algorithm show that they share common similarity in identifying readability levels—the use of Filipino stop words. Performance of other classifiers and features were also explored.

### **Development of a Filipino Speaker Diarization in Meeting Room Conversations**

***Angelica H. De La Cruz, Rodolfo C. Raga Jr***

Speaker diarization pertains to the process of determining speaker identity at a given time in an audio stream. It was first used for speech recognition and over time became useful in other applications such as video captioning and speech transcription. Recently, deep learning techniques have been applied to speaker diarization with considerable success, however, deep learning are conventionally data intensive and collecting large training samples can be difficult and expensive to collect especially for resource scarce languages. This study focused on investigating a speaker diarization approach for meeting room conversations in the Filipino language. To compensate for lack of resources, a one shot learning strategy was explored using Siamese neural network. Among the experiments conducted, the lowest diarization error rate yielded to 46%. There are, however, more parameters that can be tuned to improve the diarization results. To the best of our knowledge, no work in speaker diarization dedicated for Filipino language has yet been done.

### **Construction of Quantitative Index System of Vocabulary Difficulty in Chinese Grade Reading**

***Huiping Wang, Lijiao Yang, Huimin Xiao***

Chinese grade reading for children has a broad application prospect. In this paper, Chinese textbooks for grade 1 to 6 of primary schools published by People's Education Press are taken as data sets, and the texts are divided into 12 difficulty levels successively. The effective lexical indexes to measure the

readability of texts are discussed, and a regression model to effectively measure the lexical difficulty of Chinese texts is established. The study firstly collected 30 indexes at the text lexical level from the three dimensions of lexical richness, semantic transparency and contextual dependence, selected the 7 indexes with the highest relevance to the text difficulty through Person correlation coefficient, and finally constructed a Regression to predict the text difficulty based on Lasso Regression, ElasticNet, Ridge Regression and other algorithms. The regression results show that the model fits well, and the predicted value could explain 89.3% of the total variation of text difficulty, which proves that the quantitative index of vocabulary difficulty of Chinese text constructed in this paper is effective, and can be applied to Chinese grade reading and computer automatic grading of Chinese text difficulty.

#### **An Measurement Method of Ancient Poetry Difficulty for Adaptive Testing**

*Huiping Wang, Bihua Wang*

Chinese grade reading for children has a broad application prospect. In this paper, Chinese textbooks for grade 1 to 6 of primary schools published by People's Education Press are taken as data sets, and the texts are divided into 12 difficulty levels successively. The effective lexical indexes to measure the readability of texts are discussed, and a regression model to effectively measure the lexical difficulty of Chinese texts is established. The study firstly collected 30 indexes at the text lexical level from the three dimensions of lexical richness, semantic transparency and contextual dependence, selected the 7 indexes with the highest relevance to the text difficulty through Person correlation coefficient, and finally constructed a Regression to predict the text difficulty based on Lasso Regression, ElasticNet, Ridge Regression and other algorithms. The regression results show that the model fits well, and the predicted value could explain 89.3% of the total variation of text difficulty, which proves that the quantitative index of vocabulary difficulty of Chinese text constructed in this paper is effective, and can be applied to Chinese grade reading and computer automatic grading of Chinese text difficulty.

#### **Prosodic Realization of Focus in Changchun Mandarin and Nanjing Mandarin**

*Ying Chen, Jiajing Zhang, Bingying Ye, Chenfang Zhou*

This study was designed to explore the prosodic patterns of focus in two dialects of Mandarin. One is Changchun Mandarin and the other is Nanjing Mandarin. The current paper compares the acoustics of their prosodic realization of focus in a production experiment. Similar to standard Mandarin, which uses in-focus expansion and concomitantly post-focus compression (PFC) to code focus, results in the current study indicate that both Changchun and Nanjing speakers produced significant in-focus expansion of pitch, intensity and duration and PFC of pitch and intensity in their Mandarin dialects. Meanwhile, the results show no significant difference of prosodic changes between Changchun and Nanjing Mandarin productions. These results reveal that PFC not only exists in standard Mandarin but also in Mandarin dialects.

#### **Celebrity Profiling from Twitter Data**

*Kumar Gourav Das, Braja Gopal Patra, Sudip Kumar Naskar*

Twitter is one of the most popular social media platforms which enables users from different walks of life to have two-way communications. User categorization in Twitter platform categorizes a group of users based on their profiles, posts, and tweeting behaviors. The purpose of user categorization is to deliver relevant information to a specific class of users based on their interests. In this work, we perform user-level categorization of celebrities based on their professions. The accuracy of the proposed model is

evaluated for three different datasets(Indian,non-Indian and Combined) using various machine and deep learning frameworks. In this regard four different light weight features have been extracted(stylistic, hashtag, embeddings, and topic-based features) from the dataset and Convolutional Neural Network is the best performing model for all three datasets. The maximum accuracies obtained are 76.66%, 91.72% and 83.01% for Indian, non-Indian, and combined datasets, respectively.

**November 16, 2019**

**09:40-10:55**

**Session 5 (MT, 5)**

**Meeting Room**

**Character Decomposition for Japanese-Chinese Character-Level Neural Machine Translation**

*Jinyi Zhang, Tadahiro Matsumoto*

Twitter is one of the most popular social media platforms which enables users from different walks of life to have two-way communications. User categorization in Twitter platform categorizes a group of users based on their profiles, posts, and tweeting behaviors. The purpose of user categorization is to deliver relevant information to a specific class of users based on their interests. In this work, we perform user-level categorization of celebrities based on their professions. The accuracy of the proposed model is evaluated for three different datasets(Indian,non-Indian and Combined) using various machine and deep learning frameworks. In this regard four different light weight features have been extracted(stylistic, hashtag, embeddings, and topic-based features) from the dataset and Convolutional Neural Network is the best performing model for all three datasets. The maximum accuracies obtained are 76.66%, 91.72% and 83.01% for Indian, non-Indian, and combined datasets, respectively.

**Fusion of Image-text attention for Transformer-based Multimodal Machine Translation**

*Junteng Ma, Shihao Qin, Lan Su, Xia Li, Lixian Xiao*

In recent years, multimodal machine translation has become one of the hot research topics. In this paper, a machine translation model based on self-attention mechanism is extended for multimodal machine translation. In the model, an Image-text attention layer is added in the end of encoder layer to capture the relevant semantic information between image and text words. With this layer of attention, the model can capture the different weights between the words that is relevant to the image or appear in the image, and get a better text representation that fuses these weights, so that it can be better used for decoding of the model. Experiments are carried out on the original English-German sentence pairs of the multimodal machine translation dataset, Multi30k, and the Indonesian-Chinese sentence pairs which is manually annotated by human. The results show that our model performs better than the text-only transformer-based machine translation model and is comparable to most of the existing work, proves the effectiveness of our model.

**Statistical Machine Learning for Transliteration: Transliterating names between Sinhala, Tamil and English**

*H.S. Priyadarshani, M.D.W. Rajapaksha, M.M.S.P. Ranasinghe, K. Sarveswaran, G.V. Dias*

In this paper, we focus on building models for transliteration of personal names between the primary languages of Sri Lanka - namely Sinhala, Tamil and English. Currently, a Rule-based system has been used to transliterate names between Sinhala and Tamil. However, we found that it fails in several cases. Further, there were no systems available to transliterate names to English. In this paper, we present a hybrid approach where we use machine learning and statistical machine translation to do the transliteration. We built a parallel trilingual corpus of personal names. Then we trained a machine learner

to classify names based on the ethnicity as we found it is an influencing factor in transliteration. Then we took the transliteration as a translation problem and applied statistical machine translation to generate the most probable transliteration for personal names. The system shows very promising results compared with the existing rule-based system. It gives a BLEU score of 89 in all the test cases and produces the top BLEU score of 93.7 for Sinhala to English transliteration.

### **Confidence Modeling for Neural Machine Translation**

*Taichi Aida, Kazuhide Yamamoto*

Current methods of neural machine translation output incorrect sentences together with sentences translated correctly. Consequently, users of neural machine translation algorithms do not have a way to check which outputted sentences have been translated correctly without employing an evaluation method. Therefore, we aim to define the confidence values in neural machine translation models. We suppose that setting a threshold to limit the confidence value would allow correctly translated sentences to exceed the threshold; thus, only clearly translated sentences would be outputted. Hence, users of such a translation tool can obtain a particular level of confidence in the translation correctness. We propose some indices; sentence log-likelihood, minimum variance, and average variance. After that, we calculated the correlation between each index and bilingual evaluation score (BLEU) to investigate the appropriateness of the defined confidence indices. As a result, sentence log-likelihood and average variance calculated by probability have a weak correlation with the BLEU score. Furthermore, when we set each index as the threshold value, we could obtain high quality translated sentences instead of outputting all translated sentences which include a wide range of quality sentences like previous work.

### **Syntax-aware Transformer Encoder for Neural Machine Translation**

*Sufeng Duan, Hai Zhao, Junru Zhou, Rui Wang*

Syntax has been shown a helpful clue in various natural language processing tasks including previous statistical machine translation and recurrent neural network based machine translation. However, since the state-of-the-art neural machine translation (NMT) has to be built on the Transformer based encoder, few attempts are found on such a syntax enhancement. Thus in this paper, we explore effective ways to introduce syntax into Transformer for better machine translation. We empirically compare two ways, positional encoding and input embedding, to exploit syntactic clues from dependency tree over source sentence. Our proposed methods have a merit keeping the architecture of Transformer unchanged, thus the efficiency of Transformer can be kept. The experimental results on IWSLT' 14 German-to-English and WMT14 English-to-German show that our method can yield advanced results over strong Transformer baselines.

**11:10-12:25**

**Session 6 (SY, 5)**

**Meeting Room**

### **What affects the difficulty of Chinese syntax?**

*Yueming Du, Lijiao Yang*

The traditional measurement of sentence difficulty only focuses on lexical features but neglects syntactic features. This paper takes 800 sentences in primary school Chinese textbooks published by People's Education Press as the research object and studies their syntactic features. We use random forest to select the top five important features and then employed SVM to do the classification experiment. The precision rate, recall rate and F-score for the classification of 5 levels are respectively 50.42%, 50.40% and 50.41%, which indicates that the features we selected has practical value for the related research.

**A Chinese word segment model for energy literature based on Neural Networks with Electricity User Dictionary**

***Bochuan Song, Bo Chai, Qiang Zhang, Quanye Jia***

The traditional measurement of sentence difficulty only focuses on lexical features but neglects syntactic features. This paper takes 800 sentences in primary school Chinese textbooks published by People's Education Press as the research object and studies their syntactic features. We use random forest to select the top five important features and then employed SVM to do the classification experiment. The precision rate, recall rate and F-scored for the classification of 5 levels are respectively 50.42%, 50.40% and 50.41%, which indicates that the features we selected has practical value for the related research.

**A Study on Syntactic Complexity and Text Readability of ASEAN English News**

***Yusha Zhang, Nankai Lin, Shengyi Jiang***

The traditional measurement of sentence difficulty only focuses on lexical features but neglects syntactic features. This paper takes 800 sentences in primary school Chinese textbooks published by People's Education Press as the research object and studies their syntactic features. We use random forest to select the top five important features and then employed SVM to do the classification experiment. The precision rate, recall rate and F-scored for the classification of 5 levels are respectively 50.42%, 50.40% and 50.41%, which indicates that the features we selected has practical value for the related research.

**Tibetan word segmentation method based on CNN-BiLSTM-CRF model**

***Lili Wang, Hongwu Yang, Xiaotian Xing, Yajing Yan***

We propose a Tibetan word segmentation method based on CNN-BiLSTM-CRF model that merely uses the characters of sentence as the input so that the method does not need large-scale corpus resources and manual features for training. Firstly, we use convolution neural network to train character vectors. Then the character vectors are searched through the character lookup table to form a matrix C by stacking searched results. Then the convolution operation between the matrix C and multiple filter matrices is carried out to obtain the character-level features of each Tibetan word by maximizing the pooling. We input the character vector into the BiLSTM-CRF model, which is suitable for Tibetan word segmentation through the highway network, for getting a Tibetan word segmentation model that is optimized by using the character vector and CRF model. For Tibetan language with rich morphology, fewer parameters and faster training time make this model better than BiLSTM-CRF model in the performance of character level. The experimental results show that character input is sufficient for language modeling. The robustness of Tibetan word segmentation is improved by the model that can achieves 95.17% of the F value.

**Converting an Indonesian Constituency Treebank to the Penn Treebank Format**

***Jessica Naraiswari Arwidarasti, Ika Alfina, Adila Alfa Krisnadhi***

A constituency treebank is a key component for deep syntactic parsing of natural language sentences. For Indonesian, this task is unfortunately hindered by the fact that the only one constituency treebank publicly available is rather small with just over 1000 sentences, and not only that, it employs a format incompatible with readily available constituency treebank processing tools. In this work, we present a conversion of the existing Indonesian constituency treebank to the widely accepted Penn Treebank format. Specifically, the conversion adjusts the bracketing format for compound words as well as the POS

tagset according to the Penn Treebank format. In addition, we revised the word segmentation and POS tagging of a number of tokens. Finally, we performed an evaluation on the treebank quality by employing the Shift-Reduce parser from Stanford CoreNLP to create a parser model. A 10-fold cross-validated experiment on the parser model yields an F1-score of 70.90%.

14:00-15:00

Industry (3)

Meeting Room

Industry

14:00-15:00

Poster 2

Meeting Room

### **Duplicate Question Detection based on Neural Networks and Multi-head Attention**

*Heng Zhang, Liangyu Chen*

It is well known that using only one neural network can not get a satisfied accuracy for the problem of Duplicate Question Detection. In order to break through this dilemma, different neural networks are ensembled serially to strive for better accuracy. However, many problems, such as vanishing gradient or exploding gradient, will be encountered if the depth of neural network is blindly increased. Worse, the serial integration may be poor in computational performance since it is less parallelizable and needs more time to train. To solve these problems, we use ensemble learning with treating different neural networks as individual learners, calculating in parallel, and proposing a new voting mechanism to get better detection accuracy. In addition to the classical models based on recurrent or convolutional neural network, Multi-Head Attention is also integrated to reduce the correlation and the performance gap between different models. The experimental results in Quora question pairs dataset show that the accuracy of our method can reach 89.3%.

### **Automated Prediction of Item Difficulty in Reading Comprehension Using Long Short-Term Memory**

*Li-Huai Lin, Tao-Hsing Chang, Fu-Yuan Hsu*

Standardized tests are an important tool in education. During the test preparation process, the difficulty of each test item needs to be defined, which previously relied on expert validation or pretest for the most part, requiring a considerable amount of labor and cost. These problems can be overcome by using machines to predict the difficulty of the test items. In this study, long short-term memory (LSTM) will be used to predict the test item difficulty in reading comprehension. Experimental results show that the proposed method has a good prediction for agreement rate.

### **Improving Question Classification with Hybrid Networks**

*Yichao Cao, Miao Li, Tao Feng, Ruijing Wang, Yue Wu*

Question classification is a basic work in natural language processing, which has an important influence on question answering. Due to question sentences are complicated in many specific domains contain a large number of exclusive vocabulary, question classification becomes more difficult in these fields. To address the specific challenge, in this paper, we propose a novel hierarchical hybrid deep network for question classification. Specifically, we first take advantages of word2vec and a synonym dictionary to learn the distributed representations of words. Then, we exploit bi-directional long short-term memory networks to obtain the latent semantic representations of question sentences. Finally, we utilize convolutional neural networks to extract question sentence features and obtain the classification results by a fully-connected network. Besides, at the beginning of the model, we leverage the self-attention layer to



capture more useful features between words, such as potential relationships, etc. Experimental results show that our model outperforms common classifiers such as SVM and CNN. Our approach achieves up to 9.37% average accuracy improvements over baseline method across our agricultural dataset.

#### **Separate Answer Decoding for Multi-class Question Generation**

*Kaili Wu, Yu Hong, Mengmeng Zhu, Hongxuan Tang, Min Zhang*

Question Generation (QG) aims to automatically generate questions by understanding the semantics of source sentences and target answers. Learning to generate diverse questions for one source sentence with different target answers is important for the QG task. Despite of the success of existing state-of-the-art approaches, they are designed to merely generate a unique question for a source sentence. The diversity of answers fail to be considered in the research activities. In this paper, we present a novel QG model. It is designed to generate different questions toward a source sentence on the condition that different answers are regarded as the targets. Pointer-Generator Network(PGN) is used as the basic architecture. On the basis, a separate answer encoder is integrated into PGN to regulate the question generating process, which enables the generator to be sensitive to attentive target answers. To ease the reading, we name our model as APGN for short in the following sections of the paper. Experimental results show that APGN outperforms the state-of-the-art on SQuAD split-1 dataset. Besides, it is also proven that our model effectively improves the accuracy of question word prediction, which leads to the generation of appropriate questions.

#### **How to Answer Comparison Questions**

*Hongxuan Tang, Yu Hong, Xin Chen, Kaili Wu, Min Zhang*

"Which city has the larger population, Tokyo or New York?". To answer the question, in general, we necessarily obtain the prior knowledge about the populations of both cities, and accordingly determine the answer by numeric comparison. Using Machine Reading Comprehension (MRC) to answer such a question has become a popular research topic, which is referred to as a task of Comparison Question Answering (CQA). In this paper, we propose a novel neural CQA model which is trained to answer comparison question. The model is designed as a sophisticated neural network which performs inference in a step-by-step pipeline, including the steps of attentive entity detection (e.g., "city"), alignment of comparable attributes (e.g., "population" of the target "cities"), contrast calculation (larger or smaller), as well as binary classification of positive and negative answers. The experimentation on HotpotQA illustrates that the proposed method achieves an average F1 score of 63.09%, outperforming the baseline with about 10% F1 scores. In addition, it performs better than a series of competitive models, including DecomPRC, BERT.

#### **Coarse-to-Fine Document Ranking for Multi-Document Reading Comprehension with Answer-Completion**

*Hongyu Liu, Shumin Shi, Heyan Huang*

Multi-document machine reading comprehension (MRC) has two characteristics compared with traditional MRC: 1) many documents are irrelevant to the question; 2) the length of the answer is relatively longer. However, in existing models, not only key ranking metrics at different granularity are ignored, but also few current methods can predict the complete answer as they mainly deal with the start and end token of each answer equally. To address these issues, we propose a model that can fuse coarse-to-fine ranking processes based on document chunks to distinguish various documents more effectively.



Furthermore, we incorporate an answer-completion strategy to predict complete answers by modifying loss function. The experimental results show that our model for multi-document MRC makes a significant improvement with 7.4% and 13% respectively on Rouge-L and BLEU-4 score, in contrast with the current models on a public Chinese dataset, DuReader.

#### **Aspect-based Opinion Mining for Code-Mixed Restaurant Reviews in Indonesia**

***Andi Suciati, Indra Budi***

The goal of opinion mining is to extract the sentiment, emotions, or judgement of reviews and classified it. These reviews are very important because they can affect the decision-making from a person. In this paper, we conducted an aspect-based opinion mining research using customer reviews of restaurants in Indonesia and we focused into analyzing the code-mixed dataset. The evaluation conducted by making four scenarios namely removing stopwords without stemming, without removing stopwords but with stemming, without removing stopwords and stemming, and preprocessing with removing stopwords and stemming. We compared five algorithms which are Random Forest (RF), Multinomial Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Extra Tree classifier (ET). The models were evaluated by using 10 folds cross validation, and the results show that all aspects achieved highest scores with different algorithms. LR achieved highest score for food (81.76%) and ambience (77.29%) aspects while the highest score for price (78.71%) and service (85.07%) aspects were obtained by DT.

#### **CIEA: A Corpus for Chinese Implicit Emotion Analysis**

***Dawei Li, Jin Wang, Xuejie Zhang***

The goal of opinion mining is to extract the sentiment, emotions, or judgement of reviews and classified it. These reviews are very important because they can affect the decision-making from a person. In this paper, we conducted an aspect-based opinion mining research using customer reviews of restaurants in Indonesia and we focused into analyzing the code-mixed dataset. The evaluation conducted by making four scenarios namely removing stopwords without stemming, without removing stopwords but with stemming, without removing stopwords and stemming, and preprocessing with removing stopwords and stemming. We compared five algorithms which are Random Forest (RF), Multinomial Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Extra Tree classifier (ET). The models were evaluated by using 10 folds cross validation, and the results show that all aspects achieved highest scores with different algorithms. LR achieved highest score for food (81.76%) and ambience (77.29%) aspects while the highest score for price (78.71%) and service (85.07%) aspects were obtained by DT.

#### **An Enhancement of Malay Social Media Text Normalization for Lexicon-Based Sentiment Analysis**

***Muhammad Fakhur Razi Abu Bakar, Norisma Idris, Liyana Shuib***

Nowadays, most Malaysians use social media such as Twitter to express their opinions toward any latest issues publicly. However, user individuality and creativity of language create huge volumes of noisy words which become unsuitable as dataset for any Natural Language Processing applications such as sentiment analysis due to the irregularity of the language featured. Thus, it is important to convert these noisy words into their standard forms. Currently, there are limited studies to normalize the noisy words for Malay language. Hence, the aim of this study is to propose an enhancement of Malay social media text normalization for lexicon-based sentiment analysis. This normalizer comprises six main modules: (1) advanced tokenization, (2) Malay/English token detection, (3) lexical rules, (4) noisy token replacement, (5) n-gram, and (6) detokenization. The evaluation has been conducted and the findings show that 83.55%

achieved in Precision and 84.61% in Recall.

### **Sinhala and Tamil Speech Intent Identification From English Phoneme Based ASR**

***Yohan Karunanayake, Uthayasanker Thayasivam, Surangika Ranathunga***

Today we can find many use cases for contentbased speech classification. These include speech topic identification and spoken command recognition. Automatic Speech Recognition (ASR) sits underneath all of these applications to convert speech into textual format. However, creating an ASR system for a language is a resource-consuming task. Even though there are more than 6000 languages, all of these speech-related applications are limited to the most well-known languages such as English, because of the availability of data. There is some past research that looked into classifying speech while addressing the data scarcity. However, all of these methods have their own limitations. In this paper, we present an English language phoneme based speech intent classification methodology for Sinhala and Tamil languages. We use a pre-trained English ASR model to generate phoneme probability features and use them to identify intents of utterances expressed in Sinhala and Tamil, for which a rather small speech dataset is available. The experiment results show that the proposed method can have more than 80% accuracy for a 0.5-hour limited speech dataset in both languages.

### **Exploring Characteristics of Word Co-occurrence Network in Translated Chinese**

***Jianyu Zheng, Kun Ma, Xuemei Tang, Shichen Liang***

The translation activity involves both the source language and the target language. Compared to the standard texts in the two language, translated texts show unique language characteristics. In order to explore them from the perspective of integrality and complexity, we introduce complex network method into the study on translated Chinese. Firstly, selected the experimental texts from The ZJU Corpus of Translational Chinese(ZCTC) and its corresponding six sub-corpora, such as Press reportage and Popular lore. And then removed the punctuation and did word segmentation. Secondly, constructed a word co-occurrence network of translated Chinese. After analyzing and counting the parameters, such as their shortest path lengths, degree distributions and clustering coefficients in these networks, we verify that the word co-occurrence network of translated Chinese has small world effect and scale-free property. Finally, by constructing co-occurrence networks of standard Chinese and calculating their network parameters, we compare and verify the differences between translated Chinese and standard Chinese: "simplification and the more usage of common words. Our work expands the application of complex network in translation studies, and provides a feasible approach for studying translated Chinese based on complex networks.

### **Improving Japanese-English Bilingual Mapping of Word Embeddings based on Language Specificity**

***Yuting Song, Biligsaikhan Batjargal, Akira Maeda***

Recently, cross-lingual word embeddings have attracted a lot of attention, because they can capture semantic meaning of words across languages, which can be applied to cross-lingual tasks. Most methods learn a single mapping (e.g., a linear mapping) to transform word embeddings space from one language to another. In this paper, we propose an advanced method for improving bilingual word embeddings by adding a language-specific mapping. We focus on learning Japanese-English bilingual word embedding

mapping by considering the specificity of Japanese language. On a benchmark data set of Japanese-English bilingual lexicon induction, the proposed method achieved competitive performance compared to the method using a single mapping, with better results being found on original Japanese words.

#### **Exploring Context's Diversity to Improve Neural Language Model**

*Yanchun Zhang, Xingyuan Chen, Peng Jin, Yajun Du*

The neural language models (NLMs), such as long short term memory networks (LSTMs), have achieved great success over the years. However the NLMs usually only minimize a loss between the prediction results and the target words. In fact, the context has natural diversity, i.e. there are few words that could occur more than once in a certain length of word sequence. We report the natural diversity as context's diversity in this paper. The context's diversity, in our model, means there is a high probability that the target words predicted by any two contexts are different given a fixed input sequence. Namely the softmax results of any two contexts should be diverse. Based on this observation, we propose a new cross-entropy loss function which is used to calculate the cross-entropy loss of the softmax outputs for any two different given contexts. Adding the new cross-entropy loss, our approach could explicitly consider the context's diversity, therefore improving the model's sensitivity of prediction for every context. Based on two typical LSTM models, one is regularized by dropout while the other is not, the results of our experiment show its effectiveness on the benchmark dataset.

#### **A General Procedure for Improving Language Models in Low-Resource Speech Recognition**

*Qian Liu, Wei-Qiang Zhang, Jia Liu, Yao Liu*

It is difficult for a language model (LM) to perform well with limited in-domain transcripts in low-resource speech recognition. In this paper, we mainly summarize and extend some effective methods to make the most of the outof- domain data to improve LMs. These methods include data selection, vocabulary expansion, lexicon augmentation, multi-model fusion and so on. The methods are integrated into a systematic procedure, which proves to be effective for improving both n-gram and neural network LMs. Additionally, pre-trained word vectors using out-of-domain data are utilized to improve the performance of RNN/LSTM LMs for rescoring first-pass decoding results. Experiments on five Asian languages from Babel Build Packs show that, after improving LMs, 5.4-7.6% relative reduction of word error rate (WER) is generally achieved compared to the baseline ASR systems. For some languages, we achieve lower WER than newly published results on the same data sets.

15:15-16:30

Session 7(SE, 5)

Meeting Room

#### **Combination of Semantic Relatedness with Supervised Method for Word Sense Disambiguation**

*Qiaoli Zhou, Yuguang Meng*

We present a semi-supervised learning method that efficiently exploits semantic relatedness in order to incorporate sense knowledge into a word sense disambiguation model and to leverage system performance. We have presented sense relativeness algorithms which combine neural model learned from a generic embedding function for variable length contexts of target words on a POS-labeled text corpus, with sense-labeled data in the form of example sentences. This paper investigates the way of incorporating semantic relatedness in a word sense disambiguation setting and evaluates the method on some SensEval/SemEval lexical sample tasks. The obtained results show that such representations consistently improve the accuracy of the selective supervised WSD system.

**Levergeing Character embedding to Chinese Textual Entailment Recognition Based on Attention**

*Pengcheng Liu, Lingling Mu, Hongying Zan*

Textual Entailment Recognition is a common task in the field of Natural Language Processing, which aims to find the semantic inference relationship between two sentences. In this paper we proposed an attention based method for the task of Chinese textual entailment recognition, this method used a common framework which contains of three parts: encoding layer, attention layer, prediction layer. After training and fine-tuning, it reaches 81.52% in the test set of CCL2018 textual entailment task, which outperformed the benchmark models of the previous evaluation. Experiments also show that character embedding and CNN is useful for Chinese textual entailment recognition.

**Automatic Recognition of Allusions in Tang Poetry Based on BERT**

*Xuemei Tang, Shichen Liang, Jianyu Zheng, Renfen Hu, Zhiying Liu*

In this paper, we propose an automated method for recognize allusions in Tang poetry. The representation of text is trained by BERT pre-trained by The SiKuQuanShu. The TOP-20 candidate allusions have the highest semantic similarity to the original sentence. Then update the ranking of candidate allusions by rule-based ranking algorithm. In the final experimental results, the precision of the correct allusion same as the final ranking TOP-1 reached 63.74%, the precision of the correct allusion appears in the final ranking TOP-3 reached 70.66%, and the precision of the correct allusion appears in the final ranking TOP-5 reached 74.82%.

**Diachronic Synonymy and Polysemy: Exploring Dynamic Relation Between Forms and Meanings of Words Based on Word Embeddings**

*Shichen Liang, Jianyu Zheng, Xuemei Tang, Renfen Hu, Zhiying Liu*

In recent years, there has been a large number of publications that use distributed methods to track temporal changes in lexical semantics. However, most current researches only state the simple fact that the meaning of words has changed, lacking more detailed and in-depth analysis. We combine linguistic theory and word embedding model to study Chinese diachronic semantics. Specifically, two methods of word analogy and word similarity are associated with diachronic synonymy and diachronic polysemy respectively, and the aligned diachronic word embeddings are used to detect the changes of relationship between forms and meanings of words. Through experiments and case studies, our method achieves the ideal result. We also find that the evolution of Chinese vocabulary is closely related to social development, and there is a certain correlation between the polysemy and synonymy of the word meaning.

**Carrier Sentence Selection with Word and Context Embeddings**

*Chak Yan Yeung, John Lee, Benjamin Tsou*

This paper presents the first data-driven model for selecting carrier sentences with word and context embeddings. In computer-assisted language learning systems, fill-in-the-blank items help users review or learn new vocabulary. A crucial step in automatic generation of fill-in-the-blank items is the selection of carrier sentences that illustrate the usage and meaning of the target word. Previous approaches for carrier sentence selection have mostly relied on features related to sentence length, vocabulary difficulty and word association strength. We train a statistical classifier on a large-scale, automatically constructed corpus of sample carrier sentences for learning Chinese as a foreign language, and use it to predict the suitability of a candidate carrier sentence for a target word. Human evaluation shows that our approach leads to substantial improvement over a word co-occurrence heuristic, and that context embeddings

further enhance selection performance.

16:45-18:00

Session 8 (LR, 5)

Meeting Room

#### **A New Method of Tonal Determination for Chinese Dialects**

*Yan Li, Zhiyi Wu*

Values of the basic tones are the key to do research on dialects in China. The traditional method of determining tones by ear and the more popular method used in experimental phonetics are either inaccurate to some degree or difficult to learn. The method provided and discussed in this paper is simple and reliable, requiring the use of only Praat and fundamental frequency value. More examples are given to prove this method's effectiveness.

#### **Extremely Low Resource Text simplification with Pre-trained Transformer Language Model**

*Takumi Maruyama, Kazuhide Yamamoto*

Recent text simplification approaches regard the task as a monolingual text-to-text generation inspired by machine translation. In particular, the transformer-based translation model outperform previous methods. Although machine translation approaches need a large-scale parallel corpus, parallel corpora for text simplification are very small compared to machine translation tasks. Therefore, we attempt a simple approach which fine-tunes the pre-trained language model for text simplification with a small parallel corpus. Specifically, we conduct experiments with the following two models: transformer-based encoder-decoder model and a language model that receives a joint input of original and simplified sentences, called TransformerLM. Thus, we show that TransformerLM, which is a simple text generation model, substantially outperforms a strong baseline. In addition, we show that fine-tuned TransformerLM with only 3,000 supervised examples can achieve performance comparable to a strong baseline trained by all supervised data.

#### **Automatic Meta-evaluation of Low-Resource Machine Translation Evaluation Metrics**

*Junting Yu, Wuying Liu, Hongye He, Lin Wang*

Meta-evaluation is a method to assess machine translation (MT) evaluation metrics according to certain theories and standards. This paper addresses an automatic meta-evaluation method of machine translation evaluation based on ORANGE - Limited ORANGE, which is applied in low-resource machine translation evaluation. It is adopted when the resources are limited. And take the three n-gram-based metrics - BLEUS, ROUGE-L and ROUGE-S for experiment, which is called horizontal comparison. Also, vertical comparison is used to compare the different forms of the same evaluation metric. Compared with the traditional human method, this method can evaluate metrics automatically without extra human involvement except for a set of references. It only needs the average rank of the references, and will not be influenced by the subjective factors. And it costs less and expends less time than the traditional one. It is good for the machine translation system parameter optimization and shortens the system development period. In this paper, we use this automatic meta-evaluation method to evaluate BLEUS, ROUGE-L, ROUGE-S and their different forms based on Cilin on the Russian-Chinese dataset. The result shows the same as that of the traditional human meta-evaluation. In this way, the consistency and effectiveness of Limited ORANGE are verified.

**The Initial Research of Mongolian Literary Corpus-Take the Text of Da.Nachugdorji's Work for Instance**

*YinhuaHai*

Today, the Mongolian corpus is gradually developed from the basic resource construction stage to an indepth research covering multi-level processing or authorcorpus- based quantitative analysis, and multi-functional electronic dictionary's development. However, there are still many shortcomings and deficiencies in the collection, development and processing of literary corpus. In this paper, the author will introduces the corpus of Da.Nachugdorji's Literature and will discusses its profound significance, and fulfill multi-level processing such as lexical, syntactic and semantic annotation, as well as dissertates the preliminary processing research of Mongolian literary corpus from the perspective of statistics on the POS, word and phrase frequency and computation of lexical richness.

**Statistical Analysis of Syllable Duration of Uyghur Language**

*Askar Hamdulla, Guzalnur Dilmurat, Gulnur Arkin, Mijit Ablimit*

Phonetics is both an ancient and a young subject. Syllables are important units of speech. Based on the data requirements of speech synthesis and speech recognition, this paper studies from the perspective of experimental phonetics. Firstly, different syllable words are counted from the largescale "Speech Acoustic Parameters Database of Uyghur Language", including monosyllable words, two-syllable words, three-syllable words and four-syllable words. Secondly, the prosodic parameters are extracted, and statistical analysis is made. Accordingly, the duration distribution of different length words for male and female speakers are studied, and the fixed CV type syllable duration of consonant , the duration of vowel, the whole syllable duration and the pitch of syllable are extracted and analyzed. The effect of different vowels on the duration of CV syllables is further studied, and provided the reliable parameter basis for Uyghur speech synthesis and speech recognition.

**November 17, 2019**

**08:30-09:45**

**Session 9 (QA, 5)**

**Meeting Room**

**Examination-Style Reading Comprehension with Neural augmented Retrieval**

*Yiqing Zhang, HaiZhao, ZhuoshengZhang*

In this paper, we focus on an examination-style reading comprehension task which requires a multiple choice question solving but without a pre-given document that is supposed to contain direct evidences for answering the question. Unlike the common machine reading comprehension tasks, the concerned task requires a deep understanding into the detail-rich and semantically complex question. Such a reading comprehension task can be considered as a variant of early deep question-answering. We propose a hybrid solution to solve the problem. First, an attentive neural network to obtain the keywords in question. Then a retrieval based model is used to retrieve relative evidence in knowledge sources with the importance score of each word. The final choice is made by considering both question and evidence. Our experimental results show that our system gives stateof- the-art performance on Chinese benchmarks and shows its effectiveness on English dataset only using unstructured knowledge source.

**Automatic answer ranking based on sememe vector in KBQA**

*Yadi Li, Lingling Mu, Hao Li, Hongying Zan*

This paper proposes an answer ranking method used in Knowledge Base Question Answering (KBQA) system. This method first extracts the features of predicate sequence similarity based on sememe vector,

**IALP 2019**

predicates' edit distances, predicates' word co-occurrences and classification. Then the above features are used as inputs of the ranking learning algorithm Ranking SVM to rank the candidate answers. In this paper, the experimental results on the data set of KBQA system evaluation task in the 2016 Natural Language Processing & Chinese Computing (NLPCC 2016) show that, the method of word similarity calculation based on sememe vector has better results than the method based on word2vec. Its accuracy, recall rate and average F1 value respectively are 73.88%, 82.29% and 75.88%. The above results show that the word representation with knowledge has import effect on natural language processing.

### **Towards Robust Neural Machine Reading Comprehension via Question Paraphrases**

*Ying Li, Hongyu Li, Jing Liu*

In this paper, we focus on addressing the oversensitivity issue of neural machine reading comprehension (MRC) models. By oversensitivity, we mean that the neural MRC models give different answers to question paraphrases that are semantically equivalent. To address this issue, we first create a large-scale Chinese MRC dataset with highquality question paraphrases generated by a toolkit used in Baidu Search. Then, we quantitatively analyze the oversensitivity issue of the neural MRC models on the dataset. Intuitively, if two questions are paraphrases of each other, a robust model should give the same predictions. Based on this intuition, we propose a regularized BERT-based model to encourage the model give the same predictions to similar inputs by leveraging high-quality question paraphrases. The experimental results show that our approaches can significantly improve the robustness of a strong BERT-based MRC model and achieve improvements over the BERT-based model in terms of held-out accuracy. Specifically, the different prediction ratio (DPR) for question paraphrases of the proposed model decreases more than 10%.

### **Using WHY-type Question-Answer Pairs to Improve Implicit Causal Relation Recognition**

*Huibin Ruan, Yu Hong, Yu Sun, Yang Xu, Min Zhang*

Implicit causal relation recognition aims to identify the causal relation between a pair of arguments. It is a challenging task due to the lack of conjunctions and the shortage of labeled data. In order to improve the identification performance, we come up with an approach to expand the training dataset. On the basis of the hypothesis that there inherently exists causal relations in WHY-type Question- Answer (QA) pairs, we utilize WHY-type QA pairs for the training set expansion. In practice, we first collect WHY-type QA pairs from the Knowledge Bases (KBs) of the reading comprehension tasks, and then convert them into narrative argument pairs by Question-Statement Conversion (QSC). In order to alleviate redundancy, we use active learning (AL) to select informative samples from the synthetic argument pairs. The sampled synthetic argument pairs are added to the Penn Discourse Treebank (PDTB), and the expanded PDTB is used to retrain the neural network-based classifiers. Experiments show that our method yields a performance gain of 2.42% F1-score when AL is used, and 1.61% without using.

### **Ranking Like Human: Global-View Matching via Reinforcement Learning for Answer Selection**

*Yingxue Zhang, Ping Jian, Ruiying Geng, Yuansheng Song, Fandong Meng*

Answer Selection (AS) is of great importance for open-domain Question Answering (QA). Previous approaches typically model each pair of the question and the candidate answers independently. However, when selecting correct answers from the candidate set, the question is usually too brief to provide enough matching information for the right decision. In this paper, we propose a reinforcement learning framework that utilizes the rich overlapping information among answer candidates to help judge



the correctness of each candidate. In particular, we design a policy network, whose state aggregates both the questioncandidate matching information and the candidate-candidate matching information through a global-view encoder. Experiments on the benchmark of WikiQA and SelQA demonstrate that our RL framework substantially improves the ranking performance.

08:30-09:45

Poster 3

Meeting Room

### **A Machine Learning Model for Dating of Ancient Chinese Texts**

*Xuejin Yu, Wei Huangfu*

This paper, with the intent of solving the issues on the dating of ancient Chinese texts, takes advantage of the Long Short Term Memory Network (LSTM) to analyze and process the character sequence in ancient Chinese. In this model, each character is transformed into a high-dimensional vector, and then vectors and the non-linear relationships among them are read and analyzed by LSTM, which finally achieve the dating tags. Experimental results show that the LSTM has a strong ability to date the ancient texts, and the precision reaches about 95% in our experiments. Thus, the proposed model offers an effective method on how to date the ancient Chinese texts. It also inspires us to actively improve the time-consuming analysis tasks in the Chinese NLP field.

### **Using Mention Segmentation to Improve Event Detection with Multi-head Attention**

*Jiali Chen, Yu Hong, Jingli Zhang, Jianmin Yao*

Sentence-level event detection (ED) is a task of detecting words that describe specific types of events, including the subtasks of trigger word identification and event type classification. Previous work straightforwardly inputs a sentence into neural classification models and analyzes deep semantics of words in the sentence one by one. Relying on the semantics, probabilities of event classes can be predicted for each word, including the carefully defined ACE event classes and a "N/A" class(i.e., non-trigger word). The models achieve remarkable successes nowadays. However, our findings show that a natural sentence may possess more than one trigger word and thus entail different types of events. In particular, the closely related information of each event only lies in a unique sentence segment but has nothing to do with other segments. In order to reduce negative influences from noises in other segments, we propose to perform semantics learning for event detection only in the scope of segment instead of the whole sentence. Accordingly, we develop a novel ED method which integrates sentence segmentation into the neural event classification architecture. Bidirectional Long Short-Term Memory (Bi-LSTM) with multi-head attention is used as the classification model. Sentence segmentation is boiled down to a sequence labeling problem, where BERT is used. We combine embeddings, and use them as the input of the neural classification model. The experimental results show that the performance of our method reaches 76.8% and 74.2% F1-scores for trigger identification and event type classification, which outperforms the state-of-the-art.

### **An End-to-End Model Based on TDNN-BiGRU for Keyword Spotting**

*Shuzhou Chai, Zhenyu Yang, Changsheng Lv, Wei-Qiang Zhang*

In this paper, we proposed a neural network architecture based on Time-Delay Neural Network (TDNN)-Bidirectional Gated Recurrent Unit(BiGRU) for small footprint keyword spotting. Our model consists of three parts: TDNN, BiGRU and Attention Mechanism. TDNN models the time information and BiGRU extracts the hidden layer features of the audio. The attention mechanism generates a vector of fixed length with hidden layer features. The system generates the final score through vector linear



transformation and softmax function. We explored the step size and unit size of TDNN and two attention mechanisms. Our model has achieved a true positive rate of 99.63 %o at a 5% false positive rate.

#### **Research on New Event Detection Methods for Mongolian News**

*Shijie Wang, Feilong Bao, Guanglai Gao*

New event detection (NED) aims at detecting the first news from one or multiple streams of news stories. This paper is aimed at the field of journalism and studies the related methods of Mongolian new event detection. The paper proposes a method that combines the similarity of news content with the similarity of news elements to detect the new event. For the news content representation, according to the characteristics of the news and the different vocabulary expressions in different news categories, improve the traditional TF-IDF method. In addition, extract the main elements of the news, including time, place, subject, object, denoter, and calculate the similarity of news elements between the two news documents. Finally, the similarity between the news content and the news elements is combined to calculate the final similarity for new event detection. The experimental results show that the improved method is obvious, and the performance is significantly improved compared with the traditional new event detection system.

#### **Cross Language Information Retrieval Using Parallel Corpus with Bilingual Mapping Method**

*Rinaldi Andrian Rahmanda, Mirna Adriani, Dipta Tanaya*

New event detection (NED) aims at detecting the first news from one or multiple streams of news stories. This paper is aimed at the field of journalism and studies the related methods of Mongolian new event detection. The paper proposes a method that combines the similarity of news content with the similarity of news elements to detect the new event. For the news content representation, according to the characteristics of the news and the different vocabulary expressions in different news categories, improve the traditional TF-IDF method. In addition, extract the main elements of the news, including time, place, subject, object, denoter, and calculate the similarity of news elements between the two news documents. Finally, the similarity between the news content and the news elements is combined to calculate the final similarity for new event detection. The experimental results show that the improved method is obvious, and the performance is significantly improved compared with the traditional new event detection system.

#### **Phrase-Based Tibetan-Chinese Statistical Machine Translation**

*YONG Cuo, Xiaodong SHI, NYIMA Tashi, Yidong CHEN*

Statistical machine translation has made great progress in recent years, and Tibetan-Chinese machine translation has many needs. A phrase-based translation model is suitable for machine translation between Tibetan and Chinese, which have similar morphological changes. This paper studies the key technologies of phrase-based Tibetan-Chinese statistical machine translation, including phrasetranslation models and reordering models, and proposes a phrase-based Tibetan-Chinese statistical machine translation prototype system. The method proposed in this paper has better accuracy than Moses, the current mainstream model, in the CWMT 2013 development set, and shows great performance improvement.

#### **Design and Implementation of Burmese Speech Synthesis System Based on HMM-DNN**

*Mengyuan Liu, Jian Yang*

The research and application of speech synthesis in Chinese and English are widely used. However, most non-universal languages have relatively few electronic language resources, and speech synthesis

research is lagging behind. Burmese is a type of alphabetic writing, and Burmese belongs to Tibetan-Burmese branch of the Sino-Tibetan language. In order to develop the Burmese speech synthesis application system, this paper studies the Burmese speech waveform synthesis method, designs and implements a HMM-based Burmese speech synthesis baseline system, and based on this, introduces a deep neural network (DNN) to replace the decision tree model of HMM speech synthesis system, thereby improving the acoustic model to improve the quality of speech synthesis. The experimental results show that the baseline system is feasible, and the introduction of DNN speech synthesis system can effectively improve the quality of speech synthesis.

#### **Speech Command Classification System for Sinhala Language based on Automatic Speech Recognition**

*Thilini Dinushika, Lakshika Kavmini, Pamoda Abeyawardhana, Uthayasanker Thayasivam, Sanath Jayasena*

Conversational Artificial Intelligence is revolutionizing the world with its power of converting the conventional computer to a human-like-computer. Exploiting the speaker's intention is one of the major aspects in the field of conversational Artificial Intelligence. A significant challenge that hinders the effectiveness of identifying the speaker's intention is the lack of language resources. To address this issue, we present a domainspecific speech command classification system for Sinhala, a low-resourced language. It accomplishes intent detection for the spoken Sinhala language using Automatic Speech Recognition and Natural Language Understanding. The proposed system can be effectively utilized in value-added applications such as Sinhala speech dialog systems. The system consists of an Automatic Speech Recognition engine to convert continuous natural human voice in Sinhala language to its textual representation and a text classifier to accurately understand the user intention. We also present a novel dataset for this task, 4.15 hours of Sinhala speech corpus in the banking domain. Our new Sinhala speech command classification system provides an accuracy of 89.7% in predicting the intent of an utterance. It outperforms the state-of-the-art direct speech-to-intent classification systems developed for the Sinhala language. Moreover, the Automatic Speech Recognition engine shows the Word Error Rate as 12.04% and the Sentence Error Rate as 21.56%. In addition, our experiments provide useful insights on speech-to-intent classification to researchers in low resource spoken language understanding.

#### **Articulatory Features Based TDNN Model for Spoken Language Recognition**

*Jiawei Yu, Minghao Guo, Yanlu Xie, Jinsong Zhang*

In order to improve the performance of the Spoken Language Recognition (SLR) system, we propose an acoustic modeling framework in which the Time Delay Neural Network (TDNN) models long term dependencies between Articulatory Features (AFs). Several experiments were conducted on APSIPA 2017 Oriental Language Recognition(AP17-OLR) database. We compared the AFs based TDNN approach to the Deep Bottleneck (DBN) features based ivector and xvector systems, and the proposed approach provide a 23.10% and 12.87% relative improvement in Equal Error Rate (EER). These results indicate that the proposed approach is beneficial to the SLR task.

#### **Improved DNN-HMM English Acoustic Model Specially For Phonotactic Language Recognition**

*Wei-Wei Liu, Guo-Chun Li, Cun-Xue Zhang, Hai-Feng Yan, Jing He, Ying-Xin Gan, Yan-Miao Song, Jian-Hua Zhou, Jian-Zhong Liu, Ying Yin, Ya-Nan Li\*, Yu-Bin Huang, Ting Ruan, Wei Liu, Rui-Li Du, Hua-ying Bai, Wei Li, Sheng-Ge Zhang*

The now-acknowledged sensitive of Phonotactic Language Recognition (PLR) to the performance of the phone recognizer front-end have spawned interests to develop many methods to improve it. In this paper, improved Deep Neural Networks Hidden Markov Model (DNN-HMM) English acoustic model front-end specially for phonotactic language recognition is proposed, and series of methods like dictionary merging, phoneme splitting, phoneme clustering, state clustering and DNN-HMM acoustic modeling (DPPSD) are introduced to balance the generalization and the accusation of the speech tokenizing processing in PLR. Experiments are carried out on the database of National Institute of Standards and Technology language recognition evaluation 2009 (NIST LRE 2009). It is showed that the DPPSD English acoustic model based phonotactic language recognition system yields 2.09%, 6.60%, 19.72% for 30s, 10s, 3s in equal error rate (EER) by applying the state-of-the-art techniques, which outperforms the language recognition results on both TIMIT and CMU dictionary and other phoneme clustering methods.

### **Comparison of Prosodic Encoding of Focus in Chinese Formal and Informal Speech**

*Xin Zhong, Hui Feng*

Register refers to a mechanism which can adjust the distance between speakers and listeners. Lexical and syntactic features in registers have been investigated for a long time, but few studies have been done on acoustic features of different registers. This study tries to put focus in two different registers, namely formal and informal speech, attempting to investigate the differences of focus realization of two registers. Ten broadcasters and ten non-broadcasters are recruited to speak out 640 sentences, including 320 formal sentences and 320 informal sentences. Results show that in both formal and informal speech, the medial focus is realized with pitch expansion of on-focus syllables and compression of post focus syllables. Pause in medial focus also plays an important role in formal speech. The final focus in formal speech is realized by slight duration lengthening and pitch raising. And final focus in informal speech has longer duration of on-focus syllables. Pitch range of on-focus syllables is more expanded in informal speech than in formal speech. What's more, syllable reduction is common in informal speech rather than in formal speech.

### **A Comparative Analysis of Acoustic Characteristics between Kazak & Uyghur Mandarin Learners and Standard Mandarin Speakers**

*Gulnur Arkin, Gvljan Alijan, Askar Hamdulla, Mijit ablimit*

In this paper, based on the vowel and phonological pronunciation corpora of 20 Kazakh undergraduate Mandarin learners, 10 Uyghur learners, and 10 standard pronunciations, under the framework of the phonetic learning model and comparative analysis, the method of experimental phonetics will be applied to the Kazak and Uyghur learners. The learners and standard speaker Mandarin vowels were analyzed for acoustic characteristics, such as formant frequency values, the vowel duration similarity and other prosodic parameters were compared with the standard speaker. These results are conducive to providing learners with effective teaching-related reference information, providing reliable and correct parameters and pronunciation assessments for computer-assisted language teaching systems (CALLs), as well as improving the accuracy of multinational Chinese Putonghua speech recognition and ethnic identification.

10:00-11:15

Session 10 (AP, 5)

Meeting Room

### **Japanese grammatical simplification with simplified corpus**

*Yumeto Inaoka, Kazuhide Yamamoto*

IALP 2019

We construct a Japanese grammatical simplification corpus and established automatic simplification methods. We compare the conventional machine translation approach, our proposed method, and a hybrid method by automatic and manual evaluation. The results of the automatic evaluation show that the proposed method exhibits a lower score than the machine translation approach; however, the hybrid method garners the highest score. According to those results, the machine translation approach and proposed method present different sentences that can be simplified, while the hybrid version is effective in grammatical simplification.

#### **Learning Deep Matching-Aware Network for Text Recommendation using Clickthrough Data**

*Haonan Liu, Nankai Lin, Zitao Chen, Ke Li and Shengyi Jiang*

With the trend of information globalization, the volume of text information is exploding, which results in the information overload problem. Text recommendation system has shown to be a valuable tool to help users in such situations of information overload. In general, most researchers define text recommendation as a static problem, ignoring sequential information. In this paper, we propose a text recommendation framework with matching-aware interest extractor and dynamic interest extractor. We apply the Attention-based Long Short-Term Memory Network (LSTM) to model a user's dynamic interest. Besides, we model a user's static interest with the idea of semantic matching. We integrate dynamic interest and static interest of users' and decide whether to recommend a text. We also propose a reasonable method to construct a text recommendation dataset with clickthrough data from CCIR 2018 shared task Personal Recommendation. We test our model and other baseline models on the dataset. The experiment shows our model outperforms all the baseline models and a state-of-the-art model, and the F1-score of our model reaches 0.76.

#### **Correlational Neural Network Based Feature Adaptation in L2 Mispronunciation Detection**

*Wenwei Dong, Yanlu Xie*

Due to the difficulties of collecting and annotating second language (L2) learner's speech corpus in Computer-Assisted Pronunciation Training (CAPT), traditional mispronunciation detection framework is similar to ASR, it uses speech corpus of native speaker to train neural networks and then the framework is used to evaluate non-native speaker's pronunciation. Therefore there is a mismatch between them in channels, reading style, and speakers. In order to reduce this influence, this paper proposes a feature adaptation method using Correlational Neural Network (CorrNet). Before training the acoustic model, we use a few unannotated non-native data to adapt the native acoustic feature. The mispronunciation detection accuracy of CorrNet based method has improved 3.19% over un-normalized Fbank feature and 1.74% over bottleneck feature in Japanese speaking Chinese corpus. The results show the effectiveness of the method.

#### **Improving text simplification by corpus expansion with unsupervised learning**

*Akihiro Katsuta, Kazuhide Yamamoto*

Automatic sentence simplification aims to reduce the complexity of vocabulary and expressions in a sentence while retaining its original meaning. We constructed a simplification model that does not require a parallel corpus using an unsupervised translation model. In order to learn simplification by unsupervised manner, we show that pseudo-corpus is constructed from the web corpus and that the corpus expansion contributes to output more simplified sentences. In addition, we confirm that it is possible to learn the operation of simplification by preparing largescale pseudo data even if there is non-parallel corpus for

simplification.

# **BERT with Enhanced Layer for Assistant Diagnosis Based on Chinese Obstetric EMRs**

***Kunli Zhang, Chuang Liu, Xuemin Duan, Lijuan Zhou, Yueshu Zhao, Hongying Zan***

This paper proposes a novel method based on the language representation model called BERT (Bidirectional Encoder Representations from Transformers) for Obstetric assistant diagnosis on Chinese obstetric EMRs (Electronic Medical Records). To aggregate more information for final output, an enhanced layer is augmented to the BERT model. In particular, the enhanced layer in this paper is constructed based on strategy 1(A strategy) and/or strategy 2(A-AP strategy). The proposed method is evaluated on two datasets including Chinese Obstetric EMRs dataset and Arxiv Academic Paper Dataset (AAPD). The experimental results show that the proposed method based on BERT improves the F1 value by 19.58% and 2.71% over the stateof- the-art methods, and the proposed method based on BERT and the enhanced layer by strategy 2 improves the F1 value by 0.7% and 0.3% (strategy 1 improves the F1 value by 0.68% and 0.1%) over the method without adding enhanced layer respectively on Obstetric EMRs dataset and AAPD dataset.